CONTRIBUTION OF UNEXPLORED GENOMIC VARIANTS TO NEURODEVELOPMENTAL DISORDERS

Marcos López Sánchez

TESI DOCTORAL UPF / ANY 2017

DIRECTORS DE LA TESI

Dr. Juan R. González Ruiz

ISGLOBAL BARCELONA INSTITUTE FOR GLOBAL HEALTH

Dr. Luis A. Pérez Jurado

DEPARTAMENT OF EXPERIMENTAL AND HEALTH SCIENCES

Universitat Pompeu Fabra Barcelona

A la Vanessa i als meus pares,

Acknowledgements (Agraïments)

Primerament agrair la direcció i suport a en Dr. Juan Ramón González Ruiz i a en Dr. Luis Pérez Jurado, per donar-me la oportunitat de realitzar aquest doctorat, per el seu ampli coneixement dels seus camps de treball. Per totes les xerrades, converses i discussions que m'han ajudat comprendre millor el món de la ciència i la investigació pública, així com la tasca i responsabilitat que tenim vers la societat i vers el món. I per la seva predisposició, ja que tot i les seves responsabilitats i càrregues de treball, sempre tenen un moment per parlar i ajudar. Gràcies.

A continuació, agrair a tots els participants als estudis d'on s'han extret les dades d'aquest treball, que tot i que pertanyen a projectes internacionals i no sé si mai arribaré a conèixer a algú, sense ells no hauria sigut possible realitzar-ho. Gràcies a ells i les seves famílies és possible la investigació en aquest camp per poder identificar les causes de la patologia, el que permet cercar estratègies per tractar, pal·liar o prevenir nous casos. També a tots aquells que fan possible l'accés a aquesta informació, especialment les associacions i consorcis internacionals. Gràcies.

No puc deixar tampoc d'agrair a tots els companys de viatge durant aquests anys, especialment a aquells amb qui he compartit més moments i discussions, Carles, Carlos, Natàlia, Jose, Alejandro, Dietmar, Anna, José, Mikel, Ivon i la resta de membres del BRGE. També agrair els membres de ISGlobal, especialment a la família del CREAL: Començant per la (antiga) planta baixa, on va començar aquesta aventura, amb la Gemma, la Pati, la Montse (amb

les seves comandes de llibres), la Lourdes, la Cecilia i l'Anna (de la que no sóc admirador secret, tot i que encara hi hagi sospites). El grup dels informàtics, amb en Paco, en Jose, en Manel, en Rubén, l'Alberto, i en David; Tampoc oblidar altres treballadors, investigadors i membres del CREAL com son les Santes Gemmes, Patrones del CREAL, en Samuel, la Joana, la Mari Carmen, la Mar, la Iolanda, que tants maldecaps t'hem donat, i la Carolyn, amb qui he compartit converses molt interessants, la Ione, amb qui hem discutit de Joc de Trons fins a quedar exhausts, i la Ita, amb qui tants viatges i discussions he compartit. Donar les gràcies també a tots els membres de la sala B que no he mencionat però amb que hem compartit moltes comandes al Baluard, calendaris d'advent i molts moments que quedaran per al record, i amb tota la resta de membres de l'ISGlobal (sobretot, antic CREAL) ubicats al PRBB amb els que he tingut el plaer de compartir sort, presentacions, festes, sopars de nadal, retreats, jocs de cartes i converses. Tampoc vull oblidar-me d'agrair els grans científics que es troben davant de la institució i que es preocupen per tots, predocs inclosos, com son els doctors Manolis Kogevinas, Jordi Sunyer i Josep Mª Antó. Moltes gràcies.

Igual que la família de l'ISGlobal, també hi ha una família una mica més petita ubicada al laboratori de genètica de la UPF a la que he d'agrair les experiències i moments viscuts. Moltes gràcies a la Victoria Campuzano, amb qui tantes estones hem passat a pràctiques amb els alumnes, a la Clara, a la Ivon i la Raquel, així com la Roser Corominas, amb qui tantes converses enriquidores i maldecaps he donat, a l'Andrés, que té resposta per a tot, i tots els

membres del lab: Francesc, Paula, Eulàlia, Montse, Mar, Anna, Jairo, Eva, Marina i Mònica. Tampoc oblidar els companys i les companyes amb qui he tingut la oportunitat de coincidir: Armand, Sorina, Marta, Tina, Aïda, Gabi, Debora, María, Arturo, Guillermo, Gaudi, Marc, i tot l'equip de qGenomics. Moltes gràcies a tots.

Finalment agrair a tota la gent del meu entorn social i familiar que m'ha donat suport: Els joves de Canovelles i tots els membres del Consell de la Joventut de Canovelles, i especialment a la Maria, en Marc, la Vanessa, la Laia, en Ramon, l'Anna i l'Elena. També a tota la família, tant la de sang, com la política. Moltes i moltes gràcies a tots.

Abstract

Neurodevelopmental disorders are a group of conditions with impairments of the personal, social, academic or occupational behaviour. Autism spectrum disorder is a neurodevelopmental disorder with a high genetic component with a large fraction still unknown. In this dissertation we analyse two unexplored genomic variants: Chromosomal mosaicism and Ancestral polymorphic inversions. Chromosomal mosaic events are responsible for a small but significant proportion of patients with ASD (0.45%), with the additional detection of two loss of chromosome Y events. In addition, we developed a bioinformatic tool that improves previous methods to detect loss of chromosome Y: MADloy. In the study of ancestral polymorphic inversions, inv8p23.1 and inv17q21.31 inversions were associated with autism risk. Improvements on the method to genotype ancestral polymorphic inversions allowed the prediction of a novel inversion in 22q11.21 region which has been validated by fiber-FISH.

Resum

Els trastorns del neurodesenvolupament son un grup de condicions amb discapacitats conductuals en els àmbits personals, socials, acadèmics o ocupacionals. Els trastorns d'espectre autista són un trastorn del neurodesenvolupament amb una gran component genètica, part de la qual encara es desconeguda. En aquest treball analitzem dues variants genòmiques poc explorades: els reordenaments cromosòmics en mosaic I les inversions ancestrals polimòrfiques. Els reordenaments cromosòmics en Mosaic son responsables d'una significant però petita proporció dels pacients amb trastorn d'espectre autista (0.45%), amb la detecció addicional de dues pèrdues del cromosoma Y. Addicionalment, s'ha desenvolupat una eina bioinformàtica que millora els mètodes previs per detectar la pèrdua de cromosoma Y: MADloy. En l'estudi de les inversions ancestrals polimòrfiques, les inversions inv8p23.1 i inv17q21.31 s'han associat amb el risc d'autisme. Millores en el mètode de genotipació de les inversions ha permès la predicció de una nova inversió localitzada a la regió 22q11.21 que s'ha validat per fiber-FISH.

Preface

This dissertation is submitted for the degree of Doctor of Philosophy at the Universitat Pompeu Fabra. The research described herein was conducted under the joint supervision of Professor J. R. González and Professor L. A. Pérez-Jurado in the Institut de Salut Global (ISGlobal) and the Department of Experimental and Health Sciences, Universitat Pompeu Fabra, between November 2013 and November 2017.

This work is to the best of my knowledge original, except where acknowledgements and references are made to previous work. Neither this, nor any substantially similar dissertation has been or is being submitted for any other degree, diploma or other qualification at any other university.

Marcos López

November 2017

Contents

Page

INTRODUCTION 21

1 Neurodevelopmental disorders 21

1.1 Definition of Neurodevelopmental disorders 21

1.2 Genetic aetiology of neurodevelopmental disorders 21

1.3 Autism Spectrum Disorder 25

1.4 Genetic architecture of autism spectrum disorders 27

2 Genetic mosaicism 29

2.1 Definition of genetic mosaicism 29

2.2 Factors that affect mosaic events 32

2.3 Implications in health and disease 34

2.4 Types and mechanisms of genetic mosaicism 39

2.5 Mosaic loss of chromosome Y 44

3 Submicroscopic polymorphic inversions 46

3.1 Definition of polymorphic inversion 46

3.2 Ancestral submicroscopic polymorphic inversions 46

3.3 Implications of ancestral polymorphic inversions in health and disease 49

RATIONALE 51

OBJECTIVES 53

METHODS 55

1 Single Nucleotide polymorphisms arrays 55

2 Genomic mosaicism detection 57

3 Ancestral polymorphic inversions analysis 59

3.1 Computational analysis *59*
3.2 Experimental analysis *60*

CHAPTER 1 ..... **63**

Somatic chromosomal mosaicism: Contribution to Autism Spectrum Disorder 63

CHAPTER 2 ..... **77**

Robust estimation of mosaic loss of chromosome Y with genotype-array-intensity data 77

CHAPTER 3 ..... **85**

Nested Inversions Polymorphisms predispose chromosome 22q11.2 to meiotic rearrangements 85

DISCUSSION **93**

CONCLUSIONS **103**

REFERENCES **105**

ANNEX **121**

1 Supplementary Material CHAPTER 1 121
1.1 Supplementary Figures *121*
1.2 Supplementary Tables *126*
2 Supplementary Material CHAPTER 2 130
2.1 Online Methods *130*
2.2 Supplementary Material 1 ..... *136*
2.3 Supplementary Material 2 ..... *162*
2.4 Supplementary Material 3 ..... *173*
3 Supplementary Material CHAPTER 3 181

4 Article: Detectable clonal mosaicism in blood as a biomarker of cancer risk in Fanconi anemia 186

Figures

Figure 1: Genetic contribution to ASD population 27
Figure 2: First few weeks of embryogenesis in humans 31
Figure 3: Influence of somatic mutations in cell distribution 33
Figure 4: Confined Placental Mosaicism 37
Figure 5: DNA damaging agents and repair pathways 40
Figure 6: Chromosome Y structure 45
Figure 7: Origin of an ancestral polymorphic inversions by recombination between segmental duplication blocks 48
Figure 8: Generation of non-viable chromosomal deletions mediated by inversions in meiosis recombination 50
Figure 9: Simulation of BAF and LRR patterns for different chromosomal configurations. 56
Figure 10: triPOD graphical output 58
Figure 11: FISH Inversion assay 61
Figure 18: Structure of the 22q11.21 region in the reference genome 101

Tables

Table 1: Neurodevelopmental disorders and its genetic aetiology and prevalence 23
Table 2: Selection of disorders and diseases that show somatic mosaicism 36

INTRODUCTION

1 Neurodevelopmental disorders

1.1 Definition of Neurodevelopmental disorders

Neurodevelopmental disorders are a group of conditions with onset in the developmental period that typically manifest early in development and affect the central nervous system (Ehninger et al. 2008). The characteristics of these disorders are developmental deficits that produce impairments of the personal, social academic or occupational behaviour. Due to the affectation of the neural system, neurodevelopmental disorders have a high prevalence of co-occurring with other psychiatric disorders like mood and anxiety disorders, obsessive-compulsive disorders and schizophrenia among others (B. H. King 2016).

There are two theoretical approaches that analyse neurodevelopmental disorders, one that focuses on one or more impaired modules called *neuropsychological account* and the other that focus on cascades effects due to deficits during development called *neuroconstructivism* (D'Souza and Karmiloff-Smith 2017).

1.2 Genetic aetiology of neurodevelopmental disorders

Although multiple environmental factors acting during the prenatal, perinatal and/or early postnatal periods can cause

neurodevelopmental problems, genetics underlies the aetiology of the great majority of cases. However, the genetic aetiology of neurodevelopmental disorders is very diverse, and could be divided in a simplistic manner in four major groups as described in

Group	Condition/Disorder	Genetic Aetiology	Prevalence per 100
I (Aneuploidy)	Edwards syndrome	Trisomy of chromosome 18	0.0250
	Turner syndrome	Monosomy of chromosome X	0.0400
	Jacobs syndrome	Disomy of chromosome Y in presence of chromosome X (XYY)	0.0545
	Triple X syndrome	Trisomy of chromosome X	0.0550
	Klinefelter syndrome	Disomy of chromosome X in presence of chromosome Y (XXY)	0.0860
	Down syndrome	Trisomy of chromosome 21 (OMIM #190685)	0.1667
II (Micro-deletion or Micro-duplication)	Cri du chat syndrome	Hemizygous deletion of chromosome 5p (OMIM #123450)	0.0020
	Angelman syndrome	~4 Mb deletion (~7 genes) of chromosome 15q11-q13 (OMIM #176270 and #105830)	0.0040
	Williams syndrome	Deletion of chromosomal region 7q11.2 (OMIM #194050)	0.0044
	Prader-Willi syndrome	~4 Mb deletion (~7 genes) of chromosome 15q11-q13 (OMIM #176270 and #105830)	0.0067
	Velocardiofacial syndrome	Hemizygous deletion (1.5 to 3.0-Mb) of chromosome 22q11.2 (OMIM #188400 and #192430)	0.0250
III (Single-gene defect)	Lesch-Nyhan syndrome	Mutation in the HPRT1 gene on chromosome Xq26.2-q26.3 (OMIM #030322)	0.0005
	Lowe syndrome	Mutation in the OCRL gene on chromosome Xq26.1 (OMIM #309000)	0.0005
	Rubinstein-Taybi syndrome	Mutation in the CREBBP gene on chromosome 16p13 (OMIM #180849)	0.0008
	Cornelia de Lange syndrome	Mutation in the NIPBL gene on chromosome 15p13 (OMIM #122470)	0.0014
	Galactosemia	Homozygous or compound heterozygous mutation in the GALT gene on chromosome 9p13.3 (OMIM #230400)	0.0020
	Marfan syndrome	Mutation in the FBN1 gene on chromosome 15q21.1 (OMIM #154700)	0.0067
	Rett syndrome	Mutations in the MECP2 gene on the X-chromosome (OMIM #312750)	0.0080
	Phenylketonuria	Mutations in the PAH gene on chromosome 12q23.2 (OMIM #261600)	0.0100
	Duchenne muscular dystrophy	Mutation in the DMD gene on chromosome Xp21.2-p21.1 (OMIM #310200)	0.0143
	Tuberous sclerosis	Mutations in the TSC1 or TSC2 genes on chromosome 9q34.13 or 16p13.3 (OMIM #191100 and #613254)	0.0167
	Neurofibromatosis type 1	Mutations or deletion (~1.5 Mb) in the NF1 gene on chromosome 17q11.2 (OMIM #162200)	0.0308
	Noonan syndrome	Mutation in the PTPN11 gene on chromosome 12q24.13 (OMIM #163950)	0.0571
	Fragile X syndrome	CCG repeat expansion of the FMR1 gene (OMIM #300624)	0.0615
IV (Multifactorial)	Foetal alcohol syndrome	Multiple genes / Environmental Factors	0.1000
	Cerebral palsy		0.1500
	Tourette syndrome		0.5000
	Schizophrenia		0.5500
	Autistic spectrum disorder		1.6500
	Developmental dyscalculia		3.0000
	Attention deficit hyperactivity disorder		5.0000
	Intellectual disability		5.5000
	Developmental dyslexia		6.0000
	Developmental coordination disorder		6.5000
	Specific language impairment		7.4000
	Speech sound disorder		10.0000

(van Loo and Martens 2007) :

- Group I: Disorders characterized by aneuploidies and large chromosomal rearrangements.
- Group II: Disorders characterized by small chromosomal deletions or duplications that affect several genes.
- Group III: Disorders characterized by a mutation on one or two genes with causal implication in the disorder, either germline or mosaic.
- Group IV: Disorders with complex aetiologies that are described with a combination of environment, genetic and epigenetic factors.

The research of critical genes implicated in rare neurodevelopmental syndromes, mostly in the Groups II and III, allowed the identification of the stage of brain development and the pathways affected for correct development by studying the functionality of the genes in terms of proliferation, migration or connectivity between the neural cell populations (Engle 2010; Valiente and Marín 2010). However, complex or multifactorial disorders of group IV in which there is no clear molecular cause are more prevalent, and the proposed consequence is defects in the connectivity between cell populations caused by common mutations in several alleles (Hu, Chahrour, and Walsh 2014), what is known as the common disease/common variant model (Mitchell 2011).

Table 1: Neurodevelopmental disorders and its genetic aetiology and prevalence

Reworked from van Loo and Martens (2007) and D'Souza and Karmiloff-Smith (2017)

Group	Condition/Disorder	Genetic Aetiology	Prevalence per 100
I (Aneuploidy)	Edwards syndrome	Trisomy of chromosome 18	0.0250
	Turner syndrome	Monosomy of chromosome X	0.0400
	Jacobs syndrome	Disomy of chromosome Y in presence of chromosome X (XYY)	0.0545
	Triple X syndrome	Trisomy of chromosome X	0.0550
	Klinefelter syndrome	Disomy of chromosome X in presence of chromosome Y (XXY)	0.0860
	Down syndrome	Trisomy of chromosome 21 (OMIM #190685)	0.1667
II (Micro-deletion or Micro-duplication)	Cri du chat syndrome	Hemizygous deletion of chromosome 5p (OMIM #123450)	0.0020
	Angelman syndrome	~4 Mb deletion (~7 genes) of chromosome 15q11-q13 (OMIM #176270 and #105830)	0.0040
	Williams syndrome	Deletion of chromosomal region 7q11.2 (OMIM #194050)	0.0044
	Prader-Willi syndrome	~4 Mb deletion (~7 genes) of chromosome 15q11-q13 (OMIM #176270 and #105830)	0.0067
	Velocardiofacial syndrome	Hemizygous deletion (1.5 to 3.0-Mb) of chromosome 22q11.2 (OMIM #188400 and #192430)	0.0250
III (Single-gene defect)	Lesch-Nyhan syndrome	Mutation in the HPRT1 gene on chromosome Xq26.2-q26.3 (OMIM #030322)	0.0005
	Lowe syndrome	Mutation in the OCRL gene on chromosome Xq26.1 (OMIM #309000)	0.0005
	Rubinstein-Taybi syndrome	Mutation in the CREBBP gene on chromosome 16p13 (OMIM #180849)	0.0008
	Cornelia de Lange syndrome	Mutation in the NIPBL gene on chromosome 15p13 (OMIM #122470)	0.0014
	Galactosemia	Homozygous or compound heterozygous mutation in the GALT gene on chromosome 9p13.3 (OMIM #230400)	0.0020
	Marfan syndrome	Mutation in the FBN1 gene on chromosome 15q21.1 (OMIM #154700)	0.0067
	Rett syndrome	Mutations in the MECP2 gene on the X-chromosome (OMIM #312750)	0.0080
	Phenylketonuria	Mutations in the PAH gene on chromosome 12q23.2 (OMIM #261600)	0.0100
	Duchenne muscular dystrophy	Mutation in the DMD gene on chromosome Xp21.2-p21.1 (OMIM #310200)	0.0143
	Tuberous sclerosis	Mutations in the TSC1 or TSC2 genes on chromosome 9q34.13 or 16p13.3 (OMIM #191100 and #613254)	0.0167
	Neurofibromatosis type 1	Mutations or deletion (~1.5 Mb) in the NF1 gene on chromosome 17q11.2 (OMIM #162200)	0.0308
	Noonan syndrome	Mutation in the PTPN11 gene on chromosome 12q24.13 (OMIM #163950)	0.0571
	Fragile X syndrome	CCG repeat expansion of the FMR1 gene (OMIM #300624)	0.0615
IV (Multifactorial)	Foetal alcohol syndrome	Multiple genes / Environmental Factors	0.1000
	Cerebral palsy		0.1500
	Tourette syndrome		0.5000
	Schizophrenia		0.5500
	Autistic spectrum disorder		1.6500
	Developmental dyscalculia		3.0000
	Attention deficit hyperactivity disorder		5.0000
	Intellectual disability		5.5000
	Developmental dyslexia		6.0000
	Developmental coordination disorder		6.5000
	Specific language impairment		7.4000
	Speech sound disorder		10.0000

Something to take into consideration about neurodevelopmental genetics is that non-diseased brain is not genetically homogenous (Pack et al. 2005) with large amounts of variation derived from post-

zygotic events that include aneuploidies (30-35%) (Yurov et al. 2007), rearrangements mediated by long interspersed nuclear element-1 (L1) retrotransposon activity (Richardson, Morell, and Faulkner 2014), and other small rearrangements and point mutations. In addition, some cases of developmental brain disorders like Focal cortical dysplasia or Hemimegaloencephaly have been described to be caused by somatic mutations in critical genes, affecting only part of the brain (Poduri et al. 2013; Jamuar et al. 2014). These cases exemplify that somatic mutations have a profound impact in the neural development, and can be the source origin of complex neurodevelopmental disorders without genetic diagnostic.

1.3 Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is one of the most frequent neurodevelopmental disorders (1.47/100) (U.S. Department of Health and Human Services 2014), diagnosed in children until after age 4 years. It was previously characterized by deficits in three behavioural domains in the Diagnostic and Statistical Manual of Mental Disorders IV (DSM-IV)(Hennes and Rodes 2011), but it its new version (DSM-V) these domains have been reduced to two: social interaction and communication, and repetitive and restrictive behaviours (American Psychiatric Association 2013). There is a difference of ASD prevalence between sexes (1 in 42 for males; 1 in 189 for females), and its estimated heritability is higher than 50% (26 to 93%, depending on the more or less strict definition of the phenotype) (Miles 2011; Gaugler et al. 2014; Bourgeron 2016). The

recurrence risk in siblings of an affected individual of ASD is 7-19% (Sandin et al. 2014).

Due to its broad behavioural domains, ASD encompasses a wide range of phenotypic variation and disorders without a specific brain region or cellular type implicated. ASD diagnosis in DSM-V unifies the Autistic disorder, Asperger's disorder and pervasive-developmental disorder-not otherwise specified (PDD-NOS), and the ASD phenotypic characteristics overlap with phenotypes observed in other neuropsychiatric disorders, such as schizophrenia, attention deficit–hyperactivity disorder (ADHD) and obsessive-compulsive disorder (OCD). In addition to the shared phenotypes, the genetic component of ASD has also been shown to be shared among all these neurodevelopmental syndromes and intellectual (Mitchell 2011; Bralten et al. 2017).

One of the approaches to reduce the etiologic heterogeneity of ASD is to establish subgroups based on its familial occurrence, differentiating the idiopathic cases in families with only one individual with ASD (Simplex) from the familial cases with two or more affected individuals in 1^{st} or 2^{nd} degree relationship (Multiplex). The study of ASD patients following this differentiation reported more than threefold rate of *de novo* mutations in idiopathic cases (~7-10%) compared to multiplex cases (~2-3%) (Sebat et al. 2007; Marshall et al. 2008), but no significant differences were found in their phenotypic impairments (Oerlemans et al. 2016).

1.4 Genetic architecture of autism spectrum disorders

The genetic aetiology of ASD is very heterogeneous, with 881 genes and 2177 copy-number variants (CNV) currently associated as reported in SFARI Gene (www.gene.sfari.org, July 10, 2017), a updated database. However not a single genetic variant is present in more than 1-2% of ASD cases (Figure 1). Of those variants, cases with high penetrant variants only account for roughly 5% of the cases and are rare, while lower penetrant variants are more common and accounts for 49% of ASD inheritance in non-syndromic cases

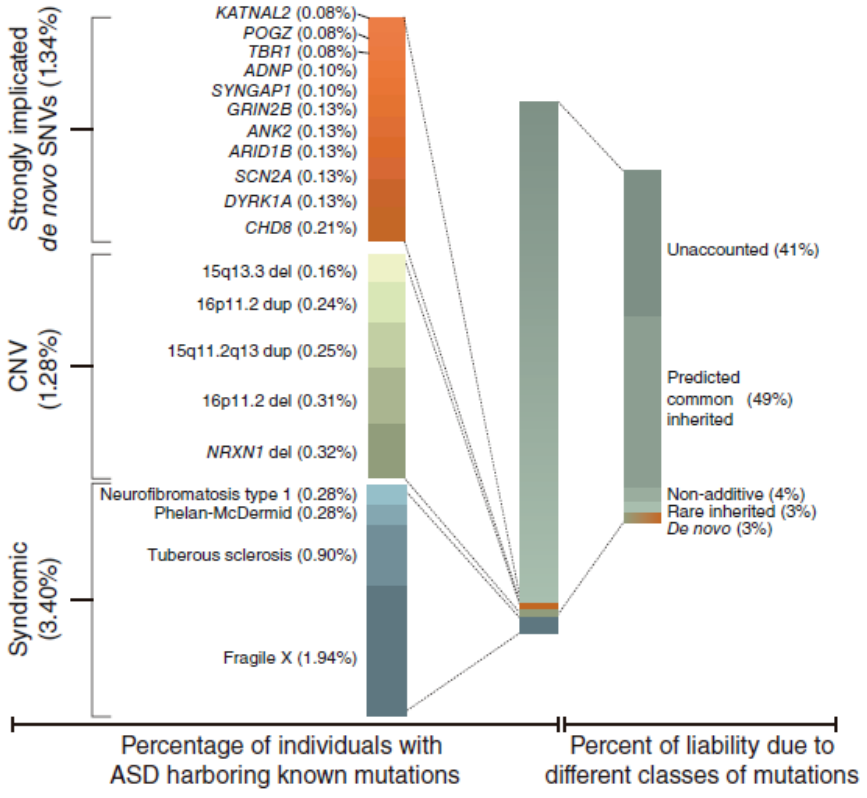


Figure 1: Genetic contribution to ASD population
Reprinted from de la Torre-Ubieta et al. (2016)

(Huguet, Ey, and Bourgeron 2013; Rosti et al. 2014; de la Torre-Ubieta et al. 2016).

This genetic heterogeneity has led to propose two genetic models not mutually exclusive and still subject to debate: A polygenic risk model (Gaugler et al. 2014), and a major gene model (O'Roak et al. 2012):

- **Polygenic risk model**: This model assumes that the combination of additive small-risk variants with the environmental factors can exceed a risk threshold for developing ASD, and is supported by the recurrence of ASD in families, the genetic heterogeneity of ASD-diagnosed siblings of ASD patients, where 69.4% of them carry different ASD-relevant mutations than its brothers (Yuen et al. 2015), and the heritability estimates in monozygotic and dizygotic twin studies.
- **Major gene model**: This model assumes that either one highly penetrant rare mutation or a limited number of moderately to highly penetrant mutations are sufficient to cause ASD, supported by the syndromic causes and the significant increase in damaging *de novo* mutations found in ASD cases as compared to their unaffected siblings.

Regarding sex bias, it has been reported that male to female prevalence ratio is lower in syndromic ASD cases (3.5:1) than in non-syndromic cases (6:1) (Elsabbagh et al. 2012), suggesting a mediation by unknown factors in common variation with protective

effect in females (E. B. Robinson et al. 2013). Furthermore, there is a significant contribution of *de novo* variants and CNVs with a paternal origin to ASD in Simplex cases (Dong et al. 2014), consistent with the effect of increased paternal age in neurodevelopmental outcomes, specifically in autism (Nybo Andersen and Urhoj 2017).

Recent studies have shown that post-zygotic mutations in mosaicism detectable in blood constitute the 7.5% of all *de novo* mutations detected in ASD, which account for a 6% of the aetiology of ASD, and that these mutations are more deleterious in ASD patient's brain-expressed critical exons than in unaffected siblings. Moreover, these mutations affect mainly genes expressed in amygdala, suggesting a possible explanation to the increased male to female ratio in non-syndromic autism (Lim et al. 2017; Dou et al. 2017).

2 Genetic mosaicism

2.1 Definition of genetic mosaicism

Genetic mosaicism is a term that describes the presence of genetically distinct cell populations in an individual due to postzygotic alterations (Youssoufian and Pyeritz 2002). It is classified depending on the affected tissues as somatic, gonadal or gonosomal mosaicism. **Somatic mosaicism** is designated when the altered mosaic cells are not carried down to the descendants because they are present only in the soma, all the cells that are not part of the

gonads. **Gonadal mosaicism** is designated when the altered mosaic cells are present only the gonads. If the mosaic variant is present in both tissues, the soma and the gonads, it is designated as **gonosomal mosaicism**.

In addition to the previous classification, mosaicism can be differentiated by the size of the mutation in two categories, **sequence mosaicism**, that includes single nucleotide variants, indels and tandem repeats, and **chromosomal mosaicism**, that includes copy-number variants, uniparental disomy and aneuploidy. The types and the mechanisms that originate these are further described in section 3.2.

Although genetic mosaicism would seem an exceptional event, it is a rule rather than exception (Fernández, Torres, and Real 2015; Abyzov et al. 2017). If we consider 10^{16} as the number of somatic cell divisions during a lifetime in an individual, a diploid human genome size of $6 \cdot 10^{9}$ base pairs, and an estimated somatic mutation rate of 10^{-8} per base pair and replicative cycle in somatic cells, we would expect an average of $10 \cdot 10^{16}$ mutations in a single individual for a lifetime (Lynch 2010; Milholland et al. 2017)). These expected mutations do not account for structural events, which have a lower rate of *de novo* generation but a length that increases its impact, and therefore the estimated number of mutations should be higher (Itsara et al. 2010). In addition, most of these mutations takes place postnatally due to the low number of divisions from zygote formation to birth (Figure 2). The generation of this diverse cell

population has been proposed to have a role in adaptation to physiological changes and cell fitness (Fernández, Torres, and Real 2015).

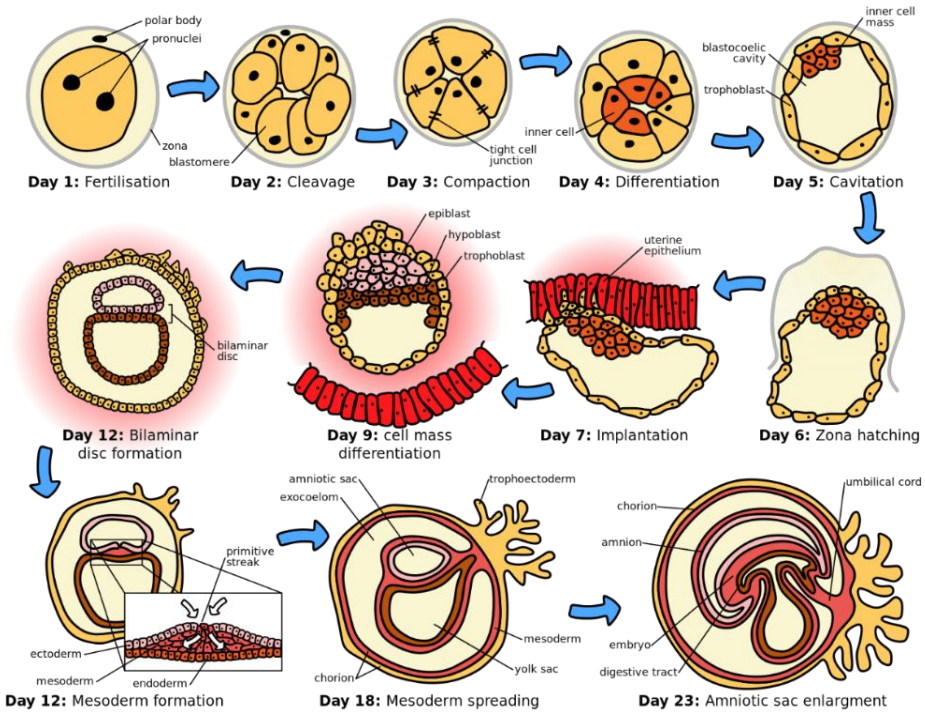


Figure 2: First few weeks of embryogenesis in humans
Made by Zephyris. CC-BY-SA-3.0 license

Mutations can arise in gonadal cells as well, where the mutation rate is higher than somatic mutation rate (10^{-11} mutations per base pair per replicative cycle) (Milholland et al. 2017), as well as structural variants (10–30% of fertilised oocytes are aneuploid, compared to only 1–2% of spermatozoa) (Hassold and Hunt 2001), but the number of divisions is much lower (401 in males, 31 in females) (Crow 2000). The consequences of mosaicism in gonads is that the alteration will be transmitted to the offspring and will be considered

a *de novo* alteration. This case is exemplified with several reported cases of inherited aneuploidies in children where the main cause is the presence of gonadal mosaicism in their parents (Delhanty 2011).

Chromosomal mosaicism detectable in blood has been associated with individual age. The prevalence of mosaic (autosomal) chromosomal events >2Mb in healthy individuals less than 50 years old is 0.23%, while it increases to 1.91% in individuals older than 75 years old (Jacobs et al. 2012). In elder male population, loss of chromosome Y detectable in peripheral blood is the most common acquired alteration during life. It has been associated with increased risk of Alzheimer (Jan P. Dumanski et al. 2016), secondary outcomes in cardiovascular defects (Haitjema et al. 2017), all-cause mortality, and nonhematologic cancer incidence (Forsberg et al. 2014).

2.2 Factors that affect mosaic events

Having into account that genetic mosaicism is a common event, its development, impact and our ability to detect it will depend in the following factors:

- **Developmental stage:** The developmental stage when the mosaic event arises will determine the tissues affected. Our whole organism derives from a unique zygote with a unique chromosomal makeup, and the following mitotic divisions in the embryo development have the potential to trigger a mutation that can be passed down to its derived cell,

generating a mosaic event that can affect several tissues (Figure 3). As developmental stage progresses, the potential of a new mutation to be passed down, affecting other tissues and generating a mosaic event will decrease.

- **Asymmetric cellular dynamics in development:** Throughout embryo development, the dynamics of the mitotic divisions are asymmetric (Ju et al. 2017). The consequences of this asymmetry are that the unequal number of divisions can lead to unbalanced proportions in cell populations even in the absence of selective pressure.

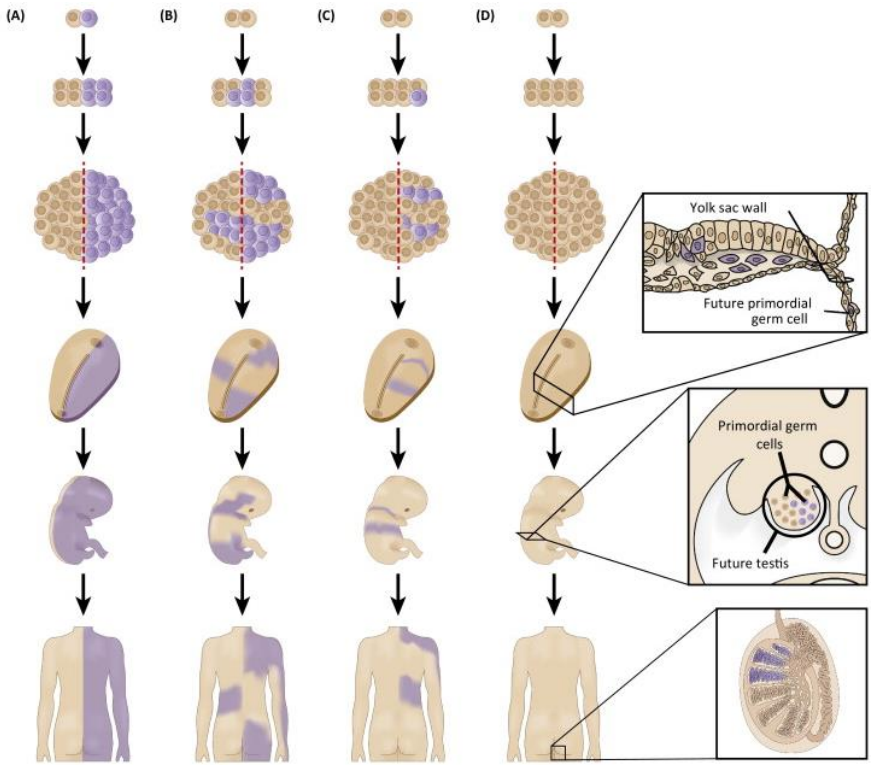


Figure 3: Influence of somatic mutations in cell distribution
Reprinted from I. Campbell, C. Shawn, P. Stankiewicz et al. (2016)

- **Selective pressure:** It can happen that the mutation makes the cell not viable, per example by affecting essential genes of the cell division process. However, when the mutation does not affect cell viability it will be subject to selective pressure, and those more suited will proliferate more, modifying the cellular proportions accordingly. There are several examples of this factor, such as mosaic monosomy/trisomy that result in foetal miscarriage for some aneuploidies if are non-mosaic.

- **Interaction between cell populations:** There are several types of interaction between mosaic cell populations described that lead to cooperation between cell populations (Burrell et al. 2013) or competition (Moreno and Rhiner 2014) that will determine the consequences of mosaicism, as well as its proliferation or removal.

- **Genes affected:** The genes affected by the mutational event will dictate the outcome of the cell population. If the gene or genes affected are essential genes required for normal cell development, the most probable consequence will be the loss of the mosaic population.

2.3 Implications in health and disease

The phenotypic consequences of genetic mosaicism are very diverse and depend on the factors stated in the previous section 2.2, as well as the expression profile of the genes and the affected tissue or

tissues. Although mosaicism is only associated with sporadic disease, there are common physiological processes that takes advantage of these events. Some examples are the adaptive immune system that promote variant generation (Di Noia and Neuberger 2007), and the polyploidy mechanisms present in human liver hepatocytes (Duncan 2013). However, the most noticeable genetic mosaic events are related with disease processes.

Regarding clinical manifestations, mosaic events can be differentiated between **mosaic manifestation of Mendelian disorders** and **disorders only manifested in mosaicism** (Table 2) (Biesecker and Spinner 2013; Gottlieb, Beitel, and Trifiro 2001). Mosaic manifestations of Mendelian disorders are mosaic forms of the same mutations that underlie disorders, inherited usually in an autosomal dominant pattern, that are compatible with life viability when constitutional. The disorder can have a milder effect in the mosaic form due to variable expressivity, exemplified in several cases of children with developmental disorders due to the arise of somatic mosaic events in early stages of development (D. A. King et al. 2015). On the other hand, disorders only manifested in mosaicism are disorders only possible in mosaic form due to incompatibility with viability when constitutional, and usually are incapable of germline transmission.

Some disorders only manifested in mosaicism are caused by the clonal proliferation of mosaic cells called aberrant clonal expansions. These expansions are the consequence of dysregulation

of genes connected to cancer development, which are frequently affected by mutational events due to the aging process, with the

Table 2: Selection of disorders and diseases that show somatic mosaicism

Extracted from Gottlieb, Beitel, and Trifiro (2001)

Mendelian disorder in mosaic	Mutated gene
Alport syndrome (OMIM # 301050)	COL4A5
Angelman syndrome (OMIM # 105830)	UBE3A
Conradi-Hunermann-Happle syndrome (OMIM # 302960)	EBP
Darier-White Disease (OMIM # 124200)	ATP2A2
Double cortex syndrome (OMIM # 300067)	DCX
Duchenne muscular dystrophy (OMIM # 310200)	DMD
Dyskeratosis congenita (OMIM # 305000)	DKC1
Fanconi's anemia	FANC gene family
Fascioscapulohumeral muscular dystrophy (OMIM # 158900)	
Friederich ataxia	Frataxin
Glycophorin A	
Haemophilia A	Factor VIII
Haemophilia B	Factor IX
Hunter disease	
Larsen syndrome	
Linear nervous sebaceous Syndrome	
Lowe syndrome	ORCI1
Marfan syndrome	FBN1
Myotonic dystrophy	
Neurofibromatosis type 1	NF1
Neurofibromatosis type 2	NF2
McCune-Albright sydnrome	
Non-McCune-Albright fibrous dysplasia	GNAS1
Pallister-Killian Syndrome	
Periventricular nodular heteropia	
Progressive osseus heteroplasia	
Proteus Syndrome	
Pseudoachrondoplasia	COMP
Retinoblastoma	RB1
Rothmund-Thompson syndrome	RECQL4
Tuberous sclerosis complex	TSC1 & TSC2
Trisomy in chromosomes 8, 9 or 14	
Von Hippel-Lindau syndrome	
X-linked hydrocephalus	

potential to develop a carcinogenic process (Forsberg, Gisselsson, and Dumanski 2016; Fernández, Torres, and Real 2015). As stated previously, one of the most common aberrant clonal expansion is the loss of chromosome Y, that will be further explained in section 2.5 because its importance in this work.

In addition to these disorders, there are also mechanisms of spontaneous genetic reversion that generate mosaic cell populations to restore pathogenic variants. Two examples are *Bloom syndrome* and *Fanconi anemia*, where intragenic recombination, back mutation and compensatory mutations mechanisms have been described in some cases in order to reverse the mutations causing the disorder (Youssoufian and Pyeritz 2002)

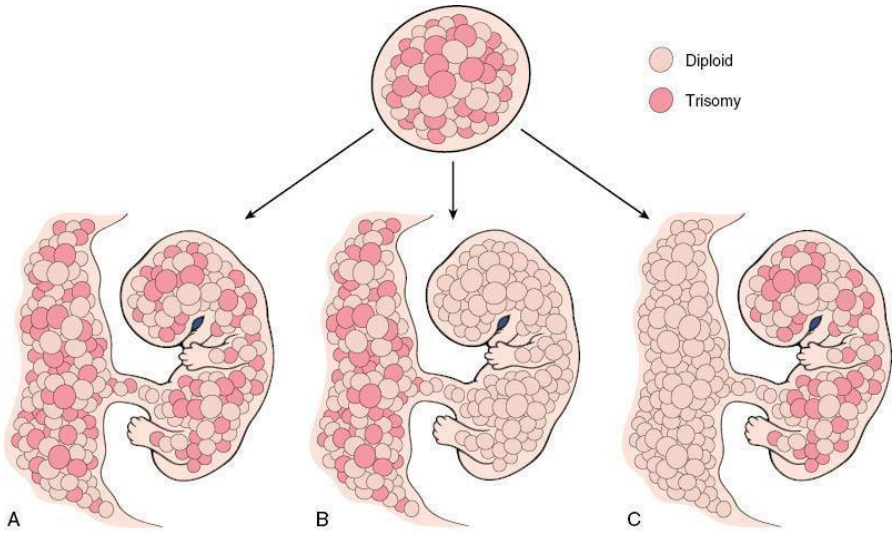


Figure 4: Confined Placental Mosaicism
A. Mosaic population of trisòmic cells in both placenta and the embryo. B. Confined placental mosaicism. C. True foetal mosaicism.
Reworked from Kalousek et al (1990)

Reversion mechanisms for aneuploidies during early development can generate mosaic populations in placental tissue, what is known as confined placental mosaicism. Confined placental mosaicism is a discrepancy between placental cells and the embryo cells genome (Figure 4). It has been reported that at least 10% of the gametes have aneuploidies (Hassold and Hunt 2001), and that up to 70% of human embryos display CNV or whole chromosomal aneuploidies in at least one blastomere during the first week of embryogenesis (Vanneste et al. 2009), but most of these anomalies do not affect the foetus thanks to a rescue mechanism that restricts the aneuploidy to the placental cells. Due to the possible consequences of aneuploidy in development, it is important during prenatal testing to differentiate confined placental mosaicism from true foetal mosaicism (Grati 2014).

Somatic mutation events also have important consequences in the analysis of transmitted alleles, recurrence risk and genetic counselling. In cases where causal disease variants are originated in gonadal mosaic events, the consideration of mathematical models that account for mosaicism can improve recurrence risk estimations (Campbell et al. 2015).

Finally, one of the main limitations when studying the mosaic events is the tissue sample analysed, as we have already commented for confined placental mosaicism. The mosaic distribution of any genetic or genomic variant is likely to be variable among the

different tissues. Thus, the information obtained with any experimental assay is always partial and depends on the cellular types included in the sample analysed. Usually, the preferred tissue for genetic analyses is peripheral blood due to its availability, easy and non-invasive extraction procedures and purification, but there are multiple examples where mosaic events were found in fibroblast DNA and not in lymphocytic DNA (Giraldo et al. 2016; Azcona, Bareille, and Stanhope 1999). Therefore, the analysed tissue, ideally more than one, must be taken in consideration when performing analysis and drawing conclusions about mosaicism.

2.4 Types and mechanisms of genetic mosaicism

When describing genetic mosaic events several types of mutations can alter the genomic sequence or the chromosome organization and generate the mosaic cell population (Figure 5). In addition to epigenetic modifications, there are six main types of mutational events that generate genetic mosaicism:

- **Sequence variants:** One of the most common types of mosaicism is due to sequence variants in the DNA. Our DNA is always exposed to endogenous and exogenous agents that damages and modify the genetic sequence. Several repair pathways exist to revert the damages generated by these agents, but in some cases these damages cannot be repaired and generates a persisting error in the sequence, in the form of base mismatch, methylation, deamination, insertion or deletion among others. Affected

cells with extensive genome instability will be removed via programmed cell death or apoptosis, but those that does not have instability can generate a mosaic population (Chatterjee and Walker 2017).

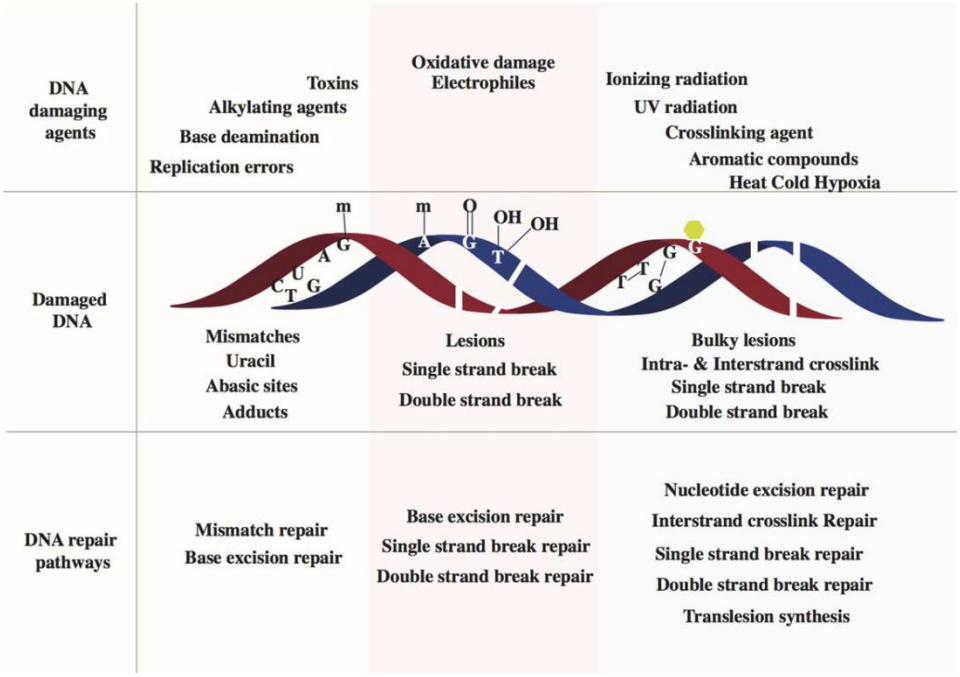


Figure 5: DNA damaging agents and repair pathways
Reprinted from Chatterjee and Walker (2017)

- **Retrotransposon-induced mutations:** Other mutational agents that generate mosaic cell populations are the short-interspersed nucleotide elements (SINE or Alu) and the long-interspersed nucleotide element-1 (LINE-1 or L1) retrotransposons. L1 retrotransposons are transposable elements present in the human genome that accounts approx. for 17% of the whole genome. Approximately 80-100 of those elements are active in an individual genome, with the

ability to retrotranspose and insert its sequence in other regions of the genome, but its functionality is restricted in somatic cells due to methylation. However, L1 retrotransposons plays an important role in neural proliferation and plasticity where they are not methylated. In addition, L1 retrotransposition is associated with large-scale rearrangements, specifically translocations (Brouha et al. 2003; Richardson, Morell, and Faulkner 2014).

- **Tandem repeat variation:** There are several expansion mechanisms of tandem repeat variation associated with known disorders (Fragile X syndrome, Huntington disease) that take place in cell division, and can generate mosaic cell populations with several degrees of the trinucleotide expansion.

- **Copy-number variants:** Regarding gene copy-number variants (CNV), two types can be differentiated, recurrent and non-recurrent. Recurrent CNVs are generated by non-homologous recombination between segmental duplications, low copy repeats sequences (>10000 bp) with high homology (>95%). Recurrent CNVs are formed during meiotic divisions and usually do not generate mosaic events. Non-recurrent CNVs are generated by replication-dependent repair pathways that involve small microhomologies (>10 bp) mainly during mitotic divisions. The main mechanisms implicated are Break-induced Replication (BIR) and

alternative nonhomologous end joining (Alt-NHEJ), but others mechanisms are also in study. (Conover and Argueso 2016; Hastings et al. 2009)

- **Uniparental Disomy:** Uniparental disomy (UPD) stretches can also be the source of a mosaic cell population. In this type of event, there is no gain or loss of a chromosomal region. Instead, the stretch has the same parental origin. If both chromosomal regions are from the same parental chromatid, it is called **isodisomy**, and if are from different chromatids of the same parental, it is called **heterodisomy**. Uniparental disomy in mosaicism can be generated through 5 mechanisms: Trisomic rescue, compensatory UPD, somatic recombination, gene conversion and mitotic nondisjunction. Trisomic rescue takes place during development and requires a disomy in one of the gametes, usually with a maternal meiotic origin. When trisomic zygote divides, one of the extra chromosomes is lost before blastocyte formation in order to be viable. The two selected chromosomes that will conform the foetus can be one from each parent, or both from the mother, generating the mosaic UPD. This process has been extensively studied in cases of confined placental mosaicism, where the samples of the chorionic villus have a trisomic karyotype and the foetus have a mosaic trisomy with large cell fractions of mosaic UPD. Compensatory UPD is a mechanism where the presence of abnormal chromosomes (ring chromosomes) or the absence of a chromosome will trigger a nondisjunction

or a duplication of the chromosome to restore the chromosomal number. UPD can also arise from somatic recombination with a reciprocal exchange between two chromatids in mitotic divisions, generating a region of isodisomy. Additionally, DNA repair mechanisms, specifically break-induced replication with homology pattern can repair damaged genes by copying the homologous strand and generating small stretches of uniparental disomy. Finally, nondisjunction in mitotic division can generate reciprocal uniparental disomies in daughter cells as well as other complex rearrangements (W. P. Robinson 2000; Lapunzina and Monk 2011; Sakofsky and Malkova 2017).

- **Aneuploidy:** Aneuploidy is the abnormality of losing or gaining one chromosome in the genome. Aneuploidy occurs in >10% of human pregnancies being a major cause of miscarriage and the major impediment in assisted reproductive technology (Nagaoka, Hassold, and Hunt 2012). It usually takes place during gamete formation in the meiotic divisions, where the main mechanisms involved are nondisjunction (NDJ), premature separation of sister chromatids (PSSC) and reverse segregation, but it can also occur in mitotic division by NDJ and Anaphase lagging (AL). NDJ can take place in meiosis I, meiosis II or mitosis, when chromosome segregation fails and homologous chromosomes (meiosis I or mitosis) or sister chromatids

(meiosis II) segregate together instead to the opposite poles. PSSC occurs in the meiosis I and implies that one of the sister chromatids segregates with the homologous chromosome. Reverse segregation also takes place in meiosis I and consists of the joint segregation of sister chromatids instead of the homologous chromosome pair to each pole, implying possible alignment problems during metaphase II. Anaphase lagging take places in mitosis due to abnormal short telomeres or mitotic checkpoint slippage. Aneuploidy in the gametes could lead to gamete complementation and the generation of constitutional uniparental disomies in the zygote (Hassold and Hunt 2001; Webster and Schuh 2017).

2.5 Mosaic loss of chromosome Y

As stated previously, mosaic loss of chromosome Y is the most common acquired genetic alteration during life detected at least in peripheral blood of males. This aberrant clonal proliferation has been associated with increased risk of Alzheimer (Jan P. Dumanski et al. 2016), secondary outcomes in cardiovascular defects (Haitjema et al. 2017), all-cause mortality, and nonhematologic cancer incidence (Forsberg et al. 2014), with some controverted results in other studies (Zhou et al. 2016).

Due to the unique characteristics of the sex chromosomes, with sex-specific and shared regions in both chromosome X and Y, these chromosomes were routinely ignored when using genome wide

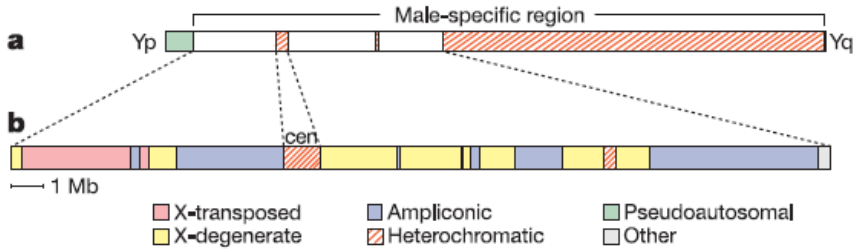


Figure 6: Chromosome Y structure
a. Representation of chromosome Y, highlighting the Male-specific region.
b. Description of the different sequence classes present in the male-specific region of chromosome Y.
Reprinted from Skaletsky et al 2003

information in association analyses (Winham, de Andrade, and Miller 2014). However, the availability of genome-wide genotypes of multiple general population samples of males has sparkled the interest of testing mosaic loss of chromosome Y (mLOY) through the study of the male-specific region of chromosome Y (msY, chrY:2,694,521-59,034,049, hg19/GRCh37) (Skaletsky et al. 2003) (Figure 6).

The current methods used to detect Loss of chromosome Y events (Forsberg et al. 2014; Jan P. Dumanski et al. 2016; Haitjema et al. 2017) rely in the median Log R Ratio of the probes in the msY region obtained from genotyping SNP arrays. The classification between mLOY and normal samples are determined by a fixed threshold that considers experimental variation of LRR values in the msY region and a confidence interval of the LRR values distribution.

3 Submicroscopic polymorphic inversions

3.1 Definition of polymorphic inversion

Inversions are copy-neutral genomic rearrangements with a change in orientation of a segment of DNA within a chromosome. These genetic variants are a common feature of the human genome (Pang et al. 2010) as well as other genomes of animals and plants, and have been reported to be implicated in speciation, population diversification, and complex diseases (Puig, Casillas, et al. 2015). Some inversions have not been fixed in humans and/or have occurred more than once during human evolution, so they are polymorphic with different frequency in the human populations.

While some large chromosomal inversions detectable by conventional cytogenetics have been known for decades, it is only in the last decade that several submicroscopic (0.1 to 5Mb in size) polymorphic inversions were detected, validated and well characterized (Kidd et al. 2008; Antonacci et al. 2009). The advances, predictions and validations in the field have led to the development of a human inversions database (Martínez-Fundichely et al. 2014). However, there are different types of polymorphic inversions, and this work will focus the attention in the study of ancestral polymorphic inversions that are present at a relevant frequency in the human population.

3.2 Ancestral submicroscopic polymorphic inversions

Ancestral submicroscopic polymorphic inversions (referred as inversions from here) are chromosomal regions with varying sizes (from 0.1 to 5 Mb) located between two breakpoints that do not result in sequence gain or loss, originated during evolution. The ancestral origin of some inversions has been thoroughly studied using sequence phylogenetic analysis and comparing chromosomal structure with human ancestors (Salm et al. 2012; González et al. 2014). The evolutionary effect of inversions is the suppression of recombination in heterozygous individuals, which enhances the conservation of the region affected by the inversion and allows the divergence between alleles (Kirkpatrick and Barton 2006).

In many cases, the ancestral polymorphic inversion breakpoints are flanked by large sequence block pairs (>100 Kb each one) of repetitive sequence called segmental duplications, with high homology between both blocks (>90%) and inverse orientation. Those inversions are thought to have arisen by non-allelic homologous recombination (NAHR) mechanisms between segmental duplication pairs (Figure 7). As previously stated, the relationship between ancestral polymorphic inversions and population diversification has important implications in the study of inverted allele frequency for populations with different ancestry background.

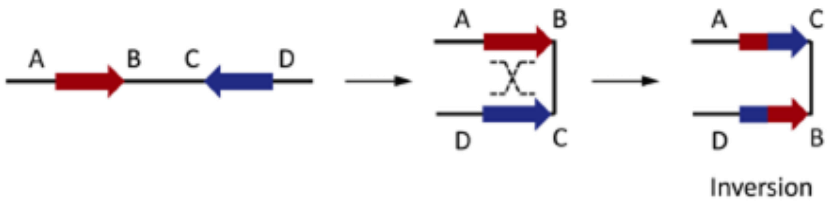


Figure 7: **Origin of an ancestral polymorphic inversions by recombination between segmental duplication blocks**

The study of these inversions poses a challenge due to the properties of segmental duplications, that hinders sequence alignment methods to identify the inversion alleles, unless one of the alleles have specific variants that allow its identification. The usual strategies to identify inverted alleles had relied on fluorescence techniques such as Fluorescence In-Situ Hybridization (FISH) and other low throughput methods (Antonacci et al. 2009), while the analysis of mutations in the inversion alleles at population level has allowed the characterization of each inversion allele and its frequency in the different human populations for a few number of inversions (Cáceres and González 2015).

However, newer sequencing methods that analyse the orientation of reads for each chromosomal strand and don't rely on the ancestral divergence between alleles have uncovered up to 111 inversions, some of which were previously detected, validated and characterized (A. D. Sanders et al. 2016)

3.3 Implications of ancestral polymorphic inversions in health and disease

The functional implications of ancestral polymorphic inversions and their possible involvement in health and disease can be due to several mechanisms alone or in combination:

- **Differential expression of genes:** Nucleotide changes in the gene sequence and/or differences in the three-dimensional structure of the different alleles, non-inverted and inverted, can alter the expression of the genes. In addition, the changes in the position or structure of regulatory elements for each allele can also affect the expression.

- **Gene disruption:** Gene sequence can be disrupted by the inverted breakpoints of the inversions, rendering the gene with reduced functionality or non-functional or generating novel fusion transcripts (Puig, Castellano, et al. 2015)

- **Predisposition to microduplications and microdeletions:** The presence of one member of a segmental duplication pair not related with the inversion breakpoints in the inversion region can predispose to generate microdeletions and microduplications by non-allelic homologous recombination in the alleles where de segmental duplication pair are in tandem (Puig, Casillas, et al. 2015).

- **Reduced fertility:** The result of a recombination event in a heterokaryotic individual for a paracentric inversion can

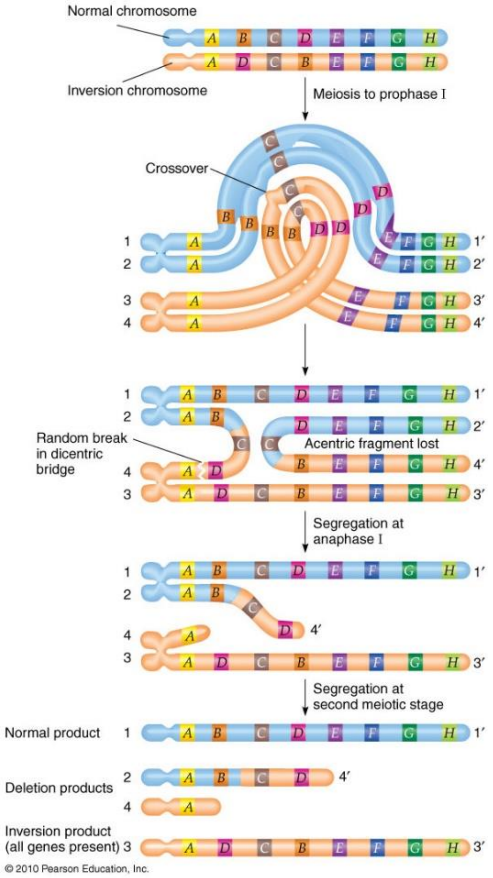


Figure 8: **Generation of non-viable chromosomal deletions mediated by inversions in meiosis recombination**

lead to the generation of dicentric or acentric chromosomes in meiosis that renders gametes non-viable (Figure 8) (Kirkpatrick 2010).

RATIONALE

The genetic aetiology of complex neurodevelopmental disorders is still incomplete. For this reason, the contribution of two unexplored genetic variants to autism spectrum disorders is analysed: Chromosomal mosaicism and ancestral polymorphic inversions.

Chromosomal mosaicism could play a relevant role in uncharacterized neurodevelopmental disorders as it has been previously implicated in some rare developmental and malformation disorders. The fact that multiple chromosomal mutations can arise in soma during embryo development and are non-inherited implies a source of variation to take into consideration.

Mosaic events affecting gonosomes could contribute in neurodevelopmental disorders. The study of sexual chromosomes is challenging, and current mosaic loss of chromosome Y analyses methods can be discussed and improved.

Ancestral polymorphic inversions can explain part of the missing common heritability in neurodevelopmental disorders. Some of these, common structural variants encompass and may alter the expression of genes important for neurodevelopment.

The study of both variants in population can be performed with SNP-array data already available of large cohorts of patients and controls, without the need to generate new genotype data.

The results we expect to obtain can give insight in the implication of chromosomal mosaic events and ancestral polymorphic inversions to neurodevelopmental disorders, as well as newer methods to detect mosaic loss of Y events that can be applied to other complex disorders with high genetic component as well.

OBJECTIVES

General Objectives

To define the contribution of two genetic variants, chromosomal mosaic events and ancestral polymorphic inversions, to some uncharacterized neurodevelopmental disorders.

Specific Objectives

The above objective will be accomplished by fulfilling the following research objectives:

- To study the contribution of chromosomal mosaic events to Autism spectrum disorders.
- To improve methods to detect mosaic loss of chromosome Y using genotyping SNP array data that enhances the methods proposed in the current state of the art.
- To study the contribution of ancestral polymorphic inversions to Autism spectrum disorders.

METHODS

1 Single Nucleotide polymorphisms arrays

SNP genotyping arrays measures common genetic variations, called single nucleotide variants (SNP), between members of a species. A SNP is a single base pair mutation at a specific site in the DNA that is usually characterized by two alleles that determines its genotype. To be classified as a SNP, a single nucleotide variant must be present in at least 1% of the population and species in which is identified.

There are two major manufacturers of SNP genotyping arrays, Illumina and Affymetrix, but both manufacturers take advantage of the principle of DNA hybridization and fluorescent dye to identify which nucleotide of the two alleles is present for a given SNP. The identification method differs between manufacturers and platforms, but both manufacturers compare the fluorescent signals of both alleles to identify the genotype. These principles also allow the identification of small insertion and deletion variants that can be present in the population. Its applications range from ancestry determination, quality control, to determination of risk alleles. Regarding the number of variants analysed in an array, up to 4.3 million of variants can be identified for a single individual for Illumina products and up to 10 million of variants for Affymetrix products can be analysed in the same experiment.

Processing the data from allele intensity to genotypes is different for each technology, with proprietary software available in order to obtain the genotypes. In fact, there is a common principle where the two signals for each SNP are normalized and processed, to finally obtain two measures, the Log R Ratio and the B Allele Frequency. The Log R Ratio (LRR) value is the normalized measure of total signal intensity and provides data relative to copy number. Since LRR is the logged ratio of the probe intensity to expected intensity, deviation from zero is evidence for copy number change. The B Allele Frequency (BAF) is derived from the ratio of allelic probe

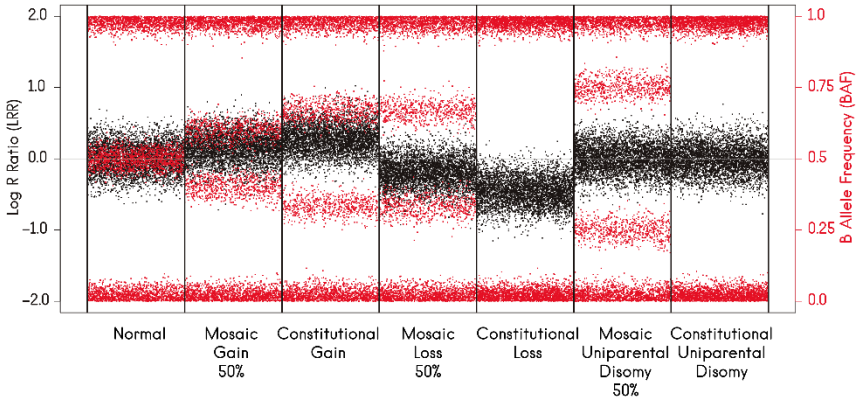


Figure 9: Simulation of BAF and LRR patterns for different chromosomal configurations.

This figure represents several chromosomal configurations in a genotyping array. Each interrogated SNP is represented by a red and a black dot. Log R Ratio points are coloured in black, while B Allele Frequency in red. The different configurations start with a normal diploid sample in the left and for each event (gain, loss, uniparental isodisomy) there is a mosaic event representation whit a 50% cellularity and a constitutional event representation.

intensity and provides data relative to allele proportion or allele composition. In normal samples the expected values are 0 and 1 for homozygous genotypes (AA, BB) and 0.5 for heterozygous genotypes (AB). (Figure 9, Normal event)

Deviations from the expectations in are indicative of a copy-number change, usually in the form of the separation of the heterozygous BAF SNPs values in two bands with the same separation from 0.5, what is known as BAF split (Figure 9). The distance between the two bands to 0.5 is known as B deviation (Bdev). Cellularity of events is calculated using B deviation (Rodríguez-Santiago et al. 2010), but this method limits the detection capacity of gain events to 14% and loss and uniparental disomy events to 7% due to B Allele Frequency background noise and the limitation in allelic ratio for gain events.

2 Genomic mosaicism detection

The detection of genomic mosaicism can be performed any method which allows to count different cells or events from the same individual. This includes from classic karyotyping analysis, to Fluorescence *in situ* hybridization (Sachdev et al. 2017), comparative-genomic hybridization and its array variation (Ostroverkhova et al. 2002; Scott et al. 2010), PCR-based methods (qPCR, DD-PCR), protein truncation test, or even sequencing methods (Whole-genome, whole-exome, targeted sequencing, single cell and molecular inversion probes among others). The resolution of the technique limits the type of genomic mosaic event that can be detected. For instance, SNP array, due to the analysis of a fixed set of variants determined by the platform, only allows to detect chromosomal mosaicism.

MAD R Package

The MAD R package was developed to detect chromosomal mosaicism using SNP array data (González et al. 2011). The method is based on the detection of consecutive positions with alterations on the B Allele Frequency with the genome alteration detection analysis (GADA) method (Pique-Regi et al. 2008). These regions are then classified according to its Log R Ratio values between gains (LRR > 0), losses (LRR < 0) and copy-neutral loss of heterozygosity (Uniparental isodisomy) events (LRR ≈ 0).

triPOD Software

triPOD Software was developed to be a more sensitive software to detect chromosomal rearrangements using SNP array data by detecting anomalies in complete trios (Baugher et al. 2013). The

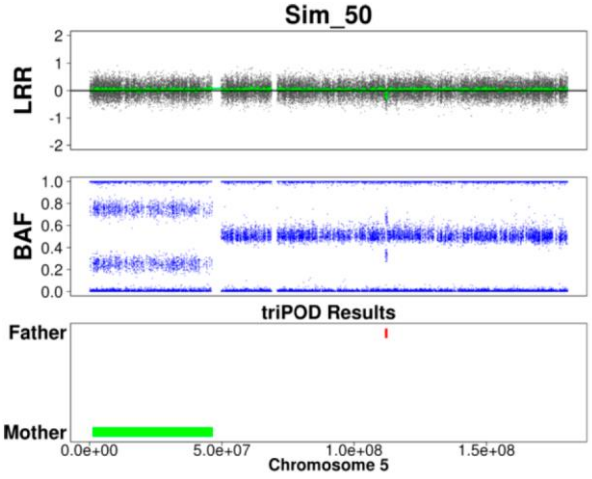


Figure 10: triPOD graphical output
Graphical representation of the LRR (upper panel; Black), BAF (mid panel; Blue) for the child of a trio where two events were detected (bottom panel): a mosaic uniparental disomy in the maternal inherited chromosome (Green) and a mosaic loss in the paternal inherited chromosome (Red) detected with triPOD.
Extracted from Baugher et al (2013)

Parent of Origin Detection method identifies which SNPs are informative for abnormal parental contribution, as well as the parental origin of the event. The advantages of this method are the detection of events at extreme cellularity values (>85% and <15%), with a higher sensitivity than MAD (Figure 10).

3 Ancestral polymorphic inversions analysis

Ancestral polymorphic inversions, as explained in section 3 of the introduction in page 46, still remain to be fully characterized in humans due to the presence of segmental duplications at the breakpoints. This section explains briefly the computational and analytical methods most relevant for this thesis, although not all the existing methods are considered.

3.1 Computational analysis

Computational analysis of inversions is performed with two R packages that use the genotype data generated with SNP arrays, but can also be used with variants obtained by other methods, like NGS data.

Detection of inversions – inveRsion R package

The inveRsion package (Cáceres et al. 2012) allows to search for inversion signals across all genome without previous knowledge of the breakpoints, by measuring differences in linkage disequilibrium between SNP blocks across inversion breakpoints for a fixed window size. A positive Bayes Information Criteria signal of the tested window indicates that the region likely harbours an inversion,

while negative signals indicates that no inversion is present (**¡Error! No se encuentra el origen de la referencia.**).

Inversion genotyping – InvClust R package

The invClust R package allows to accurately infer the inversion status on large number of subjects considering the ancestry background of each sample by relying in the differences in linkage disequilibrium between breakpoints and capturing internal haplotype groups.

3.2 Experimental analysis

Experimental analysis has been used mainly to validate candidate regions where an inversion is suspected to be present. Due to the complexity of the regions where inversions occur and the cost of the techniques (most of them only applicable to individual samples), the process is difficult and there is no gold standard.

Fluorescence *in situ* Hybridization (FISH)

Fluorescence *in situ* hybridization is a cytogenetic technique that uses fluorescent probes that bind to complementary sequences of interest in fixed chromosomes when cells are in metaphase stage. The use of different fluorescent dyes in a single assay allows testing the colour sequence of signals in order to determine its structural organisation. One example is the FISH of inv17q21.31 (Figure 11). There are several other applications of this technique other than analysis of inversions that will not be described here.

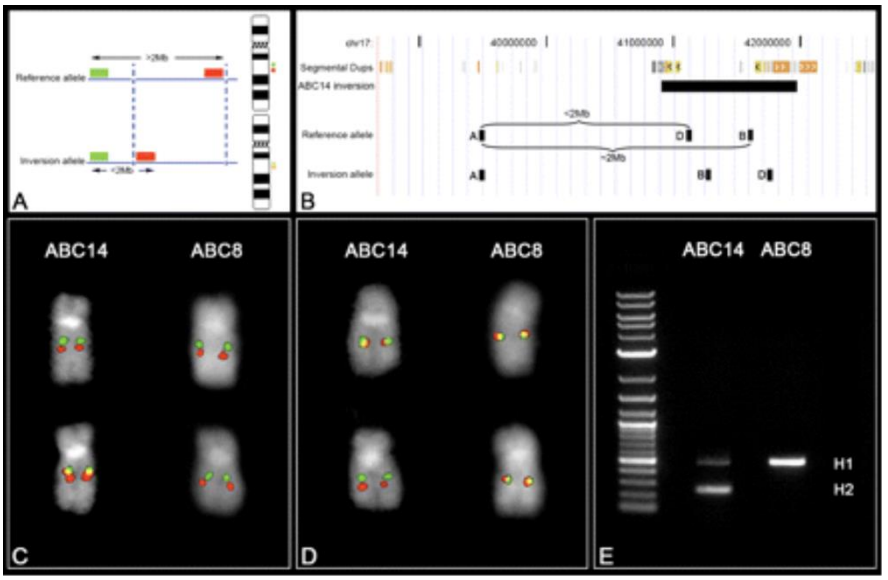


Figure 11: FISH Inversion assay

(A) Diagram of the human genomic probes. In the non-inverted state green and red labels are >2 Mb apart and appears as two distinct signals. In the inverted state, the two probes map less the 2 Mb apart and appear as a merged yellow signal. Dashed blue lines indicate inversion breakpoints. (B) FISH assay to distinguish the orientation of the 17q21.31. Probes A and B map >2 Mb apart in the non-inverted state and appear as two distinct signals (red and green). In contrast, probes A and B map less the 2 Mb apart in the inverted state and appear as a merged yellow signal (C). Reciprocal assay on the same samples using probes A and D. (E) Determination of the inversion status by PCR and SNP genotyping.

Reprinted from Antonacci et al (2009)

Fiber-FISH

Fiber-FISH is an alternative to FISH where chromosomes are attached to a slide and stretched in a straight line by applying a mechanical shear along the slide, allowing a higher resolution than conventional FISH assays.

CHAPTER 1

Somatic chromosomal mosaicism: Contribution to Autism Spectrum Disorder

Marcos López-Sánchez, Ivon Cuscó, Alejandro Cáceres, Juan R. González, Luis A. Pérez-Jurado. 2017

Submitted

Somatic chromosomal mosaicism: Contribution to Autism Spectrum Disorder

Marcos López-Sánchez, Ivon Cuscó, Alejandro Cáceres, Juan R. González, Luis A. Pérez-Jurado

Abstract

Background: Autism Spectrum Disorder (ASD) is one of the most frequent neurodevelopmental disorders. ASD has a strong genetic component in its aetiology, including de novo germ-line mutations and copy number variants. Detectable mosaicism for chromosomal rearrangements in blood has been reported in >1% of the aging population (>65y), but is very rare in young people. Somatic chromosomal mosaicism due to early developmental events could be causative of ASD.

Methods: We have studied molecular karyotypes (SNP array) from blood DNA of two large ASD datasets, the Autism Genome Project and the Simons Simplex Collection (4427 patients, 9268 parents, 2433 unaffected siblings) for chromosomal mosaic alterations >0.4Mb, using the MAD and triPOD software. As age-matched controls, we also used reported data of 5094 children with no developmental abnormalities. We independently analysed data from cell-line DNA of 564 patients and 806 parents.

Results: Chromosomal mosaic alterations (0.4-155Mb) in autosomes were detected in blood DNA of 20 out of 4427 patients (0.45%). A total of 34 events were also detected in parental samples (0.37%) and 4/7527 unaffected children (0.053%). The frequency of detectable autosomal mosaicism is significantly higher in ASD patients than in unaffected children (OR: 7.68, p=2.2·10-5), and there is no difference between parental origin of the events. Gonosomal alterations included 2 mosaic loss of chromosome Y (LOY) in two males and of chromosome X (LOX) in a female proband, while a sibling had an isoXq, one father had LOY, another a possible isoYp, and 4 mothers LOX. In cell-line DNA, autosomal frequency was similar in cases (20/564) (1 del, 3 dup, 1 tetra, 7 UPD, 1 trisx2, 7 UCT) and parents (28/806) (3 del, 4 dup, 15 UPD, 4 tris, 1 UCT, 1 tris+UCT). Regarding Unbalanced chromosomal translocations, 7 were detected in patients and just 2 in parents. UCTs are more common in patients than in parents (OR: 10.09, p=0.01).

Conclusions: Autosomal chromosomal aberrations present in mosaicism are detected in blood samples of a small but significant proportion of children with ASD (0.40%). This finding at early ages suggests that mosaicism may be present in other cell types affecting brain development and causing ASD.

Background

Autism Spectrum Disorder (ASD) is one of the most common neurodevelopmental disorders, diagnosed in children until after age 4 years, and characterized by deficits in two behavioural domains by the DSM-V: Social interaction and communication, and repetitive and restrictive behaviours [1]. The estimated population prevalence of ASD is 1 in 68 children [2], with a 4:1 male-to-female gender ratio and an estimated heritability of 50% (26 to 93%) [3–5]. There have been large genome-wide studies to identify the genetic roots of the disorder [6], looking at Genome-Wide association studies, highly penetrant alleles, mosaic single-nucleotide variants, copy-number variants and other chromosomic disorders and syndromes [7–

11], but only a 50% of the genetic component has been explained by those studies.

One of the events less explored in autism that can lead to developmental disorders is genetic mosaicism. Genetic mosaicism describes the presence of genetically distinct cell populations in an individual due to postzygotic alterations [12], and two categories can be established, variants in mosaicism and chromosomal mosaic events. Mosaic variants have an important role in the autism phenotype, contributing to the 5% of the ASD diagnoses [13–15], but the role of chromosomal mosaicism has not been explored in autism although it has an important role in unexplained developmental disorders [16]. Chromosomal mosaicism refers to events affecting large regions of the chromosome, and can be defined by single or multiple events. The origin of chromosomal mosaicism is due to aneuploidy events, events without copy-number modification, or complex events like ring-chromosomes, maker-chromosomes or translocations, detected as gain, loss and loss of heterozygosity alterations during post-zygotic divisions. There are several mechanisms implicated in the generation of chromosomal mosaicism: Non-allelic homologous recombination events, non-disjunction chromosomic events during mitosis and rescue and reversion events [12]. These alterations can be generated throughout life, and older age groups have an increased rate of mosaicism detection, exemplified by the proliferation of invasive populations of cells or non-viable constitutional chromosomal configurations that lead to diseases like neoplasia and cancer [17–19]. In addition, chromosomal mosaicism is a very rare event in general population under 14 years old, with mosaicism rates below 0.2% [20, 21]. Chromosomal mosaicism can affect one or several tissues, depending on the developmental moment when the event arose, affecting several tissues when takes place in early developmental stages [22, 23]. Considering that a mosaic event that arises in first stages of development can affect ectodermal cells, it also can potentially affect the development of the central nervous system, and thus increase the risk to be affected of ASD, because chromosomal mosaicism has been reported in cerebral cortical malformations [24]. This mechanism could have been undetected or misclassified as copy-number variants whenever is not present in all cells of the individual, leaving the affected patients undiagnosed [25].

In the present work, we quantify the frequency of early chromosomal mosaicism detected in blood for ASD by analysing SNP array data obtained from ASD affected children and compare it to the frequency in siblings from affected children and reported mosaicism rates for the same age bin. In addition, we analysed the mosaic regions for its described relationship with ASD, obtaining a small but increased mosaicism rate in probands than in siblings, closer to the mosaicism rate in parents.

Methods

Subjects

In order to study the chromosomal mosaicism rates in Autism Spectrum Disorders (ASD), two datasets comprising Trios and/or quads of affected ASD children were analysed, the Simons Simplex Collection (SSC) and the Autism Genome Project (AGP) [6] (Figure 1).

SSC cohort consists of 10220 individuals from 2591 families where each family has been characterized extensively, with a single or twin affected offspring, unaffected parents, and at least one or more unaffected siblings. All SSC DNA samples were derived from whole blood. Furthermore, DNA samples were genotyped using the Illumina Infinium Human1Mv1, Illumina Infinium Human1Mv3 Duo, or Illumina Infinium HumanOmni2.5 microarrays.

AGP cohort consists of 7880 individuals from 2614 families. A 33,65% (N = 2652) samples belongs to probands, while 52.62% (N = 4147) belongs to parents. DNA samples in AGP contains a 78% (N = 5908) derived from whole blood, a 17% (N = 1370) of samples derived from lymphoblastoid cell line, a 0.07% (N = 5) of samples originated from buccal cell samples and a 4.8% (N=371) of unknown origin. DNA samples from AGP cohort family members were genotyped using the Illumina Infinium Human1Mv1 and Human1Mv3 Duo microarray in two stages: Stage 1 and Stage 2.

APG Stage 1 consists of 4076 individuals from 1369 families. Stage 2 results include the samples from Stage 1 and consist of 7880 individuals from 2614 families in total.

Healthy siblings from the SSC cohort were analysed as control individuals to compare mosaicism rates in healthy vs affected children.

Clinical evaluation of subjects

Patients for the SSC and AGP cohorts were diagnosed using the Autism Diagnostic Interview – Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS) instruments and those patients with known karyotypic abnormalities, fragile X

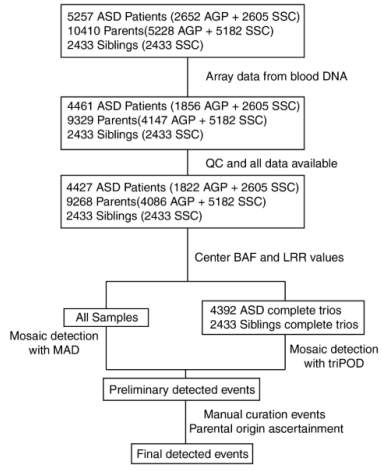


Figure 1: Detection of chromosomal mosaic events pipeline performed.

DNA from samples were analysed on high-resolution SNP array for each study. Samples with buccal or lymphoblastic DNA were not included in the main analysis. 34 samples were also removed due to a call rate < 0.95. Remaining samples were analysed for chromosomal mosaic events, individually with MAD algorithm, and in complete trios with triPOD algorithm. Detected events with both algorithms were reviewed and visually inspected. Finally, validated events were reported as results.

syndrome, trisomy 21 or other genetic disorders were excluded from the recruitment.

Probands and families with relatives who met ASD diagnostic criteria were excluded from the SSC cohort. Additionally, all Parents were screened for ASD using the Broad Autism Phenotype Questionnaire (self-report) and the Social Reciprocity Scale - Adult Research Version. In AGP not all parents were evaluated for ASD, and results were not used to exclude affected individuals and families.

In the AGP cohort, 46.2% of the families were known to be multiplex, another 38.2% were identified as simplex based on a family history indicating no known first- to third-degree relatives with ASD, and the remaining 15.6% were of unknown status, while in SSC cohort all the families are simplex. All available SSC family members that were not excluded using the criteria above were genotyped, but only parent-patient trios were genotyped for the AGP cohort even when additional siblings were available. Children were healthy individuals without neurodevelopmental disabilities. Otherwise, the whole family would be excluded from the SSC cohort.

Sample origin

The analysis requires to analyse only blood samples due to the enrichment of mosaic events in lymphoblastoid cell line samples derived from patient's blood [26]. For this reason, only microarray data originated from blood samples was considered to compare mosaicism rates. This reduced the total amount of samples to 6003 samples from AGP (1856 probands + 4147 parents) and the same number of samples from SSC.

QC and Filtering

SSC and AGP samples were not excluded on outlier levels of B Allele Frequency (BAF) values or Log R Ratio (LRR) values because these values could be due a chromosomal mosaic event. AGP Stage 1 comprised a total of 3890 samples with genotype, BAF and LRR values for 842645 probes. AGP Stage 2 comprises a total of 7745 samples with genotype, BAF and LRR values for 873678 probes. 97 samples from AGP Stage 2 with calling rates below 1% were discarded, leaving 7648 samples. Remaining samples in both AGP stages have call rates from 91 to 99.3%.

AGP Stage 1 and AGP Stage 2 were analysed independently to check for gonosomal mosaicism.

No SSC samples were removed due to bad quality, and gonosomal data were available for all the samples. Call rates in SSC vary depending on the platform, 1Mv1 (94.69-97.64%), 1Mv3 (93.41-96.11%) and Omni2.5(97.4-99.87%).

After removing the bad quality samples, 1822 probands and 4086 parents from AGP were kept for the mosaic detection comprising 4392 complete ASD trios and 2433 healthy siblings' trios, including all available siblings in each family.

Chromosomal mosaic event detection

Detection of chromosomal mosaic events (CME) was performed with two software packages, MAD [27] and triPOD [28]. Both

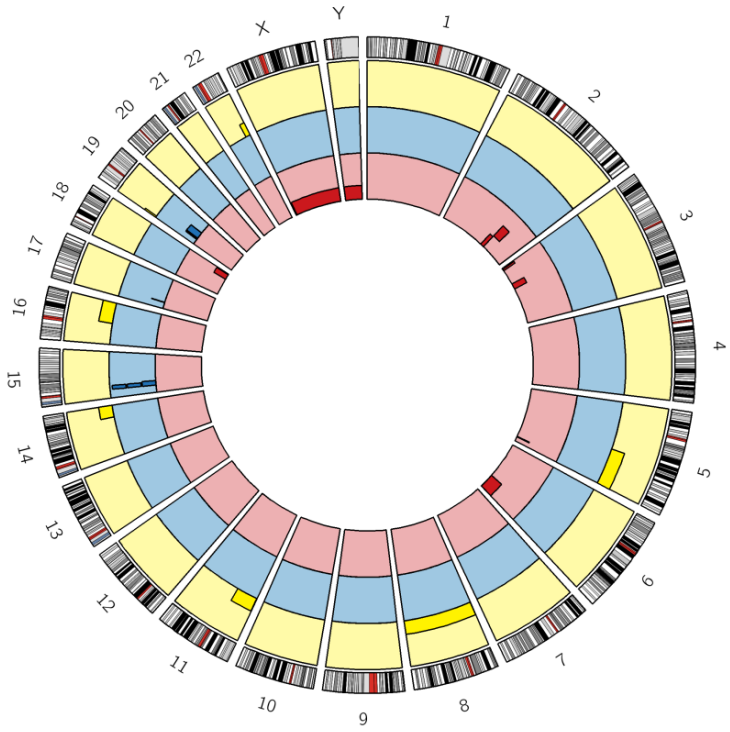


Figure 2 Circular plot of detected chromosomal mosaic events in ASD patient's blood. Yellow boxes represent LOH events, Blue boxes represent gain events and Red boxes represent loss events.

Autosomal mosaicism rates were calculated using the R software and its functions. Mosaicism rates obtained were compared against published rates in previous studies for the same age groups using the Chi-square test of independence when sample size was large and Fisher's exact test with smaller sample sizes.

Graphical representation

In order to compare the mosaic profiles of the analysed populations we performed a Circos diagram for probands with the detec ted events across all chromosomes (LOH in yellow, Gains in blue and Losses in red) [30]. Plots for each detected event was drawn with the implemented functions in MAD package.

Results

Chromosomal mosaic events >0.4 Mb detected in ASD blood samples

We analysed SNP genotyping data in search of chromosomal mosaic events larger than

Table 1: **List of chromosomal mosaic events detected in blood samples from ASD patients**
This list also includes the phenotype and other features of the patient as well as if the event has been reported previously.

		Probands' features							Detected CMEs			
Sample ID	Study	Age at Sampling	Gender	Family	Diagnosis	Verbal	Intelectual Disability	Epilepsy	Type (size in Mb)	Start (Mb)	End (Mb)	% cells	Reported (ref)
1483_7880	AGP	4 to 18	Male	Multiplex	Autism	No	NA	No	Loss 2q (24.3)	161.6	185.9	20	No
1420_2867	AGP	4 to 18	Male	Multiplex	Autism	Yes	No	No	Loss 6q (48.5)	122.2	qter	21	No
1782_7783	AGP	4 to 18	Male	Unknown	Autism	Yes	No	No	Gain 17p (0.4)	16.1	16.5	40	No
2632_7755	AGP	4 to 18	Male	Simplex	Autism	Yes	NA	No	UPD 14q (24.7)	81.6	qter	15	No
1808_5288	AGP	4 to 18	Male	Simplex	Autism	No	NA	No	UPD 22q (9.9)	39.7	qter	20	Yes (1)(4)
11671.p1	SSC	7	Male	Simplex	Autism	Yes	No	Yes	Loss 2q (11.7)	150.0	161.6	33	No
13362.p1	SSC	4	Male	Simplex	Autism	Yes	No	No	Loss 3p (6.2)	3.1	9.3	33	No
11238.p1	SSC	14	Male	Simplex	Autism-ASD	Yes	Yes	No	Loss 5q (1.7)	164.7	166.5	31	No
12246.p1	SSC	16	Male	Simplex	Autism	Yes	No	No	Loss 18q (17.7)	55.5	73.2	17	No
14687.p1	SSC	6	Male	Simplex	Autism	Yes	Yes	No	Gain 15q (12.0) *	cen	30.3	70	Yes (5)
13006.p1	SSC	10	Male	Simplex	Autism	Yes	Yes	No	UPD 8 (146.2)	pter	qter	74	Yes (3)
11679.p1	SSC	6	Male	Simplex	Autism	Yes	No	No	UPD 11p (45.8)	pter	46.6	5	No
14466.p1	SSC	8	Male	Simplex	Autism	No	Yes	No	UPD 16q (43.8)	45.0	qter	12	No
12245.p1	SSC	6	Female	Simplex	Autism	Yes	No	Yes	UPD 19p (3.5)	pter	3.7	10	No
13603.p1	SSC	4	Male	Simplex	Autism-ASD	Yes	Yes	No	Loss 9p (0.21) †	30.0	30.2	37	No
12007.p1	SSC	9	Female	Simplex	Autism	Yes	Yes	Yes	Gain 15q (8.5) *	cen	26.8	73	Yes (2)(5)
									UPD 5q (78.8)	102.0	qter	5	No
11270.p1	SSC	6	Male	Simplex	Autism	No	Yes	No	Loss 3p (18.7)	60.1	78.8	28	No
									Loss 8p (0.12) †	12.8	12.9	27	No
14556.p1	SSC	8	Male	Simplex	Autism	Yes	No	No	Gain 19p (3.0) *	14.8	17.8	20	No
									Gain 19p (16.5) *	18.4	34.9	20	No
2300_7693	AGP	4 to 18	Female	Unknown	Autism	Yes	Yes	No	Monosomy X (LOX)	pter	qter	23	Yes (1)(4)
642_47	AGP	4 to 18	Male	Multiplex	Autism	No	NA	No	Monosomy X (LOY)	pter	qter	85	No
2477_6778	AGP	4 to 18	Male	Unknown	Autism	Yes	No	No	Monosomy X (LOY) #	pter	qter	69	No

LOX Loss of X, LOY Loss of Y
* Complex rearrangement, possible marker
^Reported constitutional duplication of 76kb within region
(1) Pinto et al 2010
(2) Sanders et al 2011
(3) Pinto et al 2014
(4) Sanders et al 2015

0.4Mb. DNA was obtained from peripheral blood samples a data belongs to two ASD trio-based studies: The Autism Genome Project Study [6] (AGP) (N = 1822) and the Simons Simplex Collection Dataset [9] (SSC) (N = 2605). Of the 4427 samples belonging to ASD probands analysed, we found an autosomal mosaicism rate of 0.43% (N = 19, CI95 = 0.26 to 0.67) and a gonosomal mosaicism rate of 0.07% (N = 3, CI95 = 0.014 to 0.020). The autosomal mosaic events detected includes 7 (41.2%) copy loss events, 1 (5.9%) copy gain events, 6 (35.3%) UPD events, 1 (5.9%) marker gain

Table 2: Summary of the genes affected by the chromosomal mosaic events detected
Number of genes, number of autism-associated genes and support of the associated genes in each region of the chromosomal mosaic events detected in autism spectrum disorder patients as reported in the SFARI gene database.

Chromosome	CME	Event and region	# genes in region	# autism-associated genes	Gene Support by type
2	loss	del 2q24.2-q32.1	118	19	Functional:2; Genetic Association:8; Rare Single Gene variant:8; Rare Single Gene variant, Genetic Association:1
2	loss	del 2q23.2-q24.2	37	2	Rare Single Gene variant:2
3	loss	del 3p14.2-p12.3	42	9	Functional:2; Genetic Association:2; Rare Single Gene variant:4; Rare single gene variant/multigenic CNV:1
3	loss	del 3p25.2-p26.3	16	2	Genetic Association:2
5	UPD†	UPD 5q21.1-qter	507	31	Functional:1; Genetic Association:8; Multigenic CNV:1; Rare Single Gene variant:20; Syndromic:1
5	loss	del 5q34	0	0	
6	loss	del 6q22.1-qter	238	13	Functional:2; Genetic Association:3; Rare Single Gene variant:7; Syndromic:1
8	UPD	UPD 8	720	23	Functional:1; Genetic association:2; Genetic association/functional:1; Rare Single Gene variant:14; Rare Single Gene variant, Genetic Association:1; Syndromic:4
11	UPD	UPD 11pter-p11.2	391	11	Functional:3; Genetic Association:3; Multigenic CNV:3; Rare single Gene variant:2
14	UPD	UPD 14q31.1-qter	337	4	Rare Single Gene variant:4
15	gain	dup 15cen-q13.1	101	10	Genetic Association:3; Multigenic CNV:4; Rare Single Gene variant:2; Syndromic:1
15	gain*†	dup 15cen-q13.1	101	10	Genetic Association:3; Multigenic CNV:4; Rare Single Gene variant:2; Syndromic:1
15	gain*	dup 15cen-15q13.3	116	15	Functional:1; Genetic Association:3; Multigenic CNV:4; Rare Single Gene variant:6; Syndromic:1
16	UPD	UPD 16q11.2-qter	368	12	Functional:1; Genetic association:1; Multigenic CNV:1; Rare Single Gene variant:8; Rare Single Gene variant/genetic association:1
17	gain	dup 17p11.2	12	0	
18	loss	del 18q21.32-q23	50	4	Functional:2; Rare Single Gene variant:2
19	gain*‡	dup 19p13.12-p13.11	81	1	Rare single Gene variant:1
19	gain*‡	dup 19p13.11-q12	83	0	
19	UPD	UPD 19pter-p13.3	124	4	Functional:1; Rare Single Gene variant:2; Rare Single Gene variant, Genetic Association:1
22	UPD	UPD 22q13.2-qter	122	7	Functional:1; Rare Single Gene variant:5; Syndromic:1
X	Monosomy X	Loss of X	16 +	1	Rare Single Gene variant, Genetic Association:1
Y	Monosomy X	Loss of Y	16 +	1	Rare Single Gene variant, Genetic Association:1

UPD Uniparental disomy, CNV Copy-number variant
* Complex rearrangement, possible marker
† Multiple events in patient 12007.p1
‡ Complex event in patient 14556.p1
+ Genes in PAR regions and/or escaping X-inactivation

event, 2 (11.7%) multiple/complex events (1 UPD + marker gain event, 1 double marker gain event), while gonosomal events includes 2 Loss of Y events and 1 Loss of X event. Clonality values of the detected events ranges from 5 to 85%. Meanwhile, 14% (N = 3) of the detected events are reported gains of the 15q region, with no additional recursive events detected (Figure 2, Table 1, Document S1). No differences in parental origin of the chromosomal mosaic events were found (Paternal = 11, Maternal = 9), but the two copy-gain events detected in chromosome 15 have a maternal origin.

We further explored the SNP data on DNA from peripheral blood from unaffected patient's siblings of the same age bin belonging to the SSC study (N = 2433). We considered those samples as healthy controls to compare mosaicism rates. Of the samples analysed, we found an autosomal mosaicism rate of 0.12% (N = 3, CI_{95} = 0.03 to 0.39), and a gonosomal mosaicism rate of 0.01% (N = 1, CI_{95} = 0.0002 to 0.03). Additionally, we analysed the SNP data on DNA from

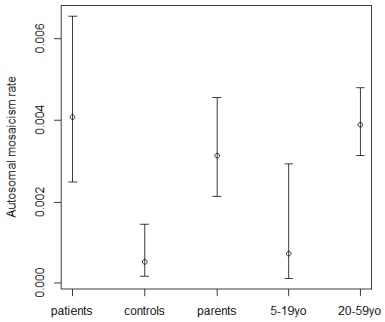


Figure 3: Autosomal mosaicism rates detected in blood

Autosomal mosaicism rates in blood for the analysed groups and the reported values for the same age groups in other studies

peripheral blood from patient's parents from both datasets (N = 9268). Of those samples, we detected an autosomal mosaicism rate of 0.31% (N = 29, CI_{95} = 0.21 to 0.46), and a gonosomal mosaicism rate of 0.08% (N = 7, CI_{95} = 0.03 to 0.16). Description of those events is available in Supplementary Table 2 and images of each detected rearrangement in Supplementary Figures 1.

In order to increase statistical power, we added the results obtained by King et al [16] after analysing autosomal chromosomal mosaic events larger than 2Mb in 5094 children of the same age bin unaffected by developmental disorders. Those samples belong to blood and saliva samples from two longitudinal studies, Twin Early Development Study and Avon Longitudinal Study of Parents and Children. The resulting total number of controls samples is 7527, and the autosomal mosaicism rate in controls after adding the detected event is 0.05% (N = 4, CI_{95} = 0.017 to 0.146). Next, autosomal mosaicism rates were compared between the different groups. Although autosomal mosaicism rate in patients was significantly higher than in controls ($p = 9.798 \cdot 10\text{-}6$, OR = 8.11, CI95 = 2.69 to 32.77), it was not significantly different from parents as we would expect due to differences in age ($p = 0.2824$, OR = 1.4, CI95 = 0.73 to 2.54). In addition, we compared mosaicism rates in parents and controls against reported mosaicism rates in general population for the corresponding age bins, but no significant differences were found (Figure 3).

The analysis of the chromosomal mosaic events regions in patients has yield a high number of autism-associated genes, as reported by the SFARI gene database, with 19 (90%) of the patients with autism-associated genes in the chromosomal mosaic event region (Table 2). There are 2 (10%) patients without genes in the chromosomal mosaic event region which are patient 1782_7783 and patient 11238.p1. Patient 1782_7783 has a 0.4 Mb gain in 17p11.2 region encompassing CENPV, UBB and TRPV2 and part of the PIGL genes. This region is included in the Yuan-Harel-Lupski syndrome (OMIM 616652), the Potocki-Lupski syndrome (OMIM 610883) and the CHIME syndrome (OMIM 605947), related with mental retardation and Autism. On the other hand, patient 11238.p1 has a 1.7Mb loss in the 5q34 region, but no genes or disorders are described in this region.

<u>Chromosomal mosaic events <0.4 Mb detected in ASD blood samples</u>

In addition to the larger chromosomal mosaic events, two smaller mosaic deletion events (<0.4Mb) were detected in probands from the SSC study in the 8p22 and 9p21 regions, with a size of 120 and 210Kb respectively. The region deleted in the 8p22 includes the whole C8orf79 (TRM9L) gene that is expressed mainly in cerebellum. The region deleted in 9p21 does not affect directly any gene, but the closest gene in the region is the LINC9 gene that is also mainly expressed in brain tissue.

Chromosomal mosaic events detected in ASD lymphoid cell line samples

In addition to the analysis of blood samples, we explored independently the SNP array data from lymphoid cell lines for AGP dataset belonging to patients (N = 564) and parents (N = 806). The events detected in patients shows that there is a 4% (N 23, CI95 = 2.6 to 6.1) of autosomal mosaicism rate, and no gonosomal mosaic events were detected. The detected events include 2 (8.7%) copy loss events, 3 (13.6%) copy gain events, 9 (39.1%) UPD events, 7 (30.4%), 1 (4.3%) marker gain event, 1 (4.3%) multiple event (double whole chromosome gain event) and 7 (30.4%) multiple/complex events (6 unbalanced chromosomal translocations and 1 whole chromosome gain event + unbalanced chromosomal translocation).

Regarding parental samples, we found a 4% (N = 32, CI95 = 2.8 to 5.6) of autosomal mosaicism rate and a 1.9% (N = 15, CI95 = 1.1 to 3.1) of gonadal mosaicism rate. Detected events includes 6 (13.3%) copy loss events, 4 (8.9%) copy gain events, 15 (33.3%) UPD events, 4 (8.9%) whole chromosome gain events, 2 (4.4%) multiple/complex events (1 unbalanced chromosomal translocation and 1 Unbalanced chromosomal translocation + whole chromosome gain event), 1 (2.2%) autosomic/gonosomic event (Whole chromosome gain X + copy loss event) and 13 (28.9%) gonadal events.

Although autosomal mosaicism rates in LCL were not comparable against blood, there are no significative differences between mosaicism rates in patients' vs parental LCL samples. However, there is a very significative increase in the number of unbalanced chromosomal translocation events in mosaic affecting probands (N = 7) than in parents (N = 2) (p = 0.037, OR = 5.04, CI95 0.95 to 49.95). Moreover, 57.1% (N = 4, CI95 = 20.2 to 88.2) of the unbalanced chromosomal translocations in probands affects chromosome 9.

Discussion

Chromosomal mosaic events are large structural variants presents only in a subpopulation of cells of an organism that have been linked to aging, aberrant clonal event proliferation and developmental disorders [16, 23, 31, 32]. The high prevalence of ASD and low genetic component explained so far pushed to explore other genetic variants. In the present work we analysed the contribution of chromosomal mosaic events detected in blood to autism spectrum disorders.

Although the detected autosomal chromosomal mosaic event rate is very low in ASD probands' blood, (0.43%), it is significant higher compared to healthy children of the same age bin ($p = 9.798 \cdot 10^{-6}$, OR 8.11, CI95 2.69 to 32.77 - Fisher exact test) Due to the lower size threshold used (>0.4Mb) than the ones reported in previous analysis (>2Mb), this autosomal chromosomal mosaic rate is not comparable to the rates detected in developmental disorders[16, 33] and general population[23, 31, 32]. However, the ability to detect these events open a door to establish the genetic cause of the ASD phenotype in unresolved cases. Some of the ASD cases detected with autosomal or gonosomal mosaicism have not been previously reported (N = 15, Table 1, Supplementary table 1), only marker chromosome 15 cases (N 2), a case with chromosome X monosomy, a case with a whole chromosome 8 UPD and a case with 22q terminal UPD were reported when analyzing the same dataset for copy-number alterations [8, 9, 11, 34] (Supplementary table 4). In addition, the gene content of the affected regions correlates with epilepsy phenotype in some patients, which agrees with the results obtained by Lamar [24] where they establish a relationship between mosaic mutations detected in blood and cortical malformations generated by those mutations. In addition, only two ASD probands of the 23 individuals with chromosomal mosaic events have previously reported sequence variants associated with autism (Supplementary table 5), that are not in the mosaic altered region, and that could be causal of the autistic traits.

Following this relationship, the chromosomal mosaic events were detected in blood samples from probands, the cell fractions detected and the early ages of the patients (4-18yo) suggest a prenatal origin, with other tissues other than mesodermal tissues (blood) affected. However, no sample from other tissues was analysed for these patients to check if ectodermal tissues were affected. The chromosomal mosaic alterations detected also must consider that genetic mosaicism has been reported in brain as a common event [35–37], and can enhance the phenotypic effect of the reported mosaic events.

Regarding the parental origin of the autosomal mosaic events, when we consider all the detected events together, there are no significative difference between parental origin rates. However, the distribution between the different types of events don't show any bias in parental origin for deletions (Paternal = 3, Maternal = 3), but there is a significative bias in maternal origin for duplications (Paternal = 0, Maternal = 3), consistent with the reported results in copy-number variants transmission associated with intellectual disability, developmental delay and congenital anomalies [38].

In the analysis of the chromosomal mosaic events in lymphoblastoid cell lines, the results obtained are in concordance with the higher chromosomal mosaic rate reported in blood when comparing versus the healthy siblings. However, the known effects of

lymphoblastoid transformation process in the generation of chromosomal variation does not allows to establish a relationship between the events and the phenotype. One of the questions that arises is the possibility that the proliferating clones are present in the patients' blood sample, which is consistent with the hypothesis that genetic mosaicism is very common, with several different variations present in the soma [18], and the ones with the chromosomal rearrangement in mosaic detected have been positively selected in the transformation process. In addition, the detection of translocations not detected in whole blood samples, together with the high number of translocations affecting chromosome 9, seems to indicate a susceptibility of chromosome 9 to be affected by LCL transformation, or that maybe that somatic rearrangements are present in blood.

Conclusions

In conclusion, autosomal chromosomal aberrations in mosaicism detected in blood samples of children with ASD is small but significant (0.40%). The finding of those events at early ages suggest the presence of the mosaic event in other cell types with a different germ layer origin, affecting brain development and causing ASD.

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 2013. doi:10.1176/appi.books.9780890425596.744053.

2. U.S. Department of Health and Human Services. Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010. MMWR Surveill Summ. 2014;63:1–21. doi:24670961.

3. Miles JH. Autism spectrum disorders—A genetics review. Genet Med. 2011;13:278–94. doi:10.1097/GIM.0b013e3181ff67ba.

4. Bourgeron T. Current knowledge on the genetics of autism and propositions for future research. C R Biol. 2016;339:300–7. doi:10.1016/j.crvi.2016.05.004.

5. Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, et al. Most genetic risk for autism resides with common variation. Nat Genet. 2014;46:881–5. doi:10.1038/ng.3039.

6. Hu-Lince D, Craig DW, Huentelman MJ, Stephan DA. The Autism Genome Project: Goals and strategies. Am J PharmacoGenomics. 2005;5:233–46. doi:10.2165/00129785-200505040-00004.

7. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, et al. Structural Variation of Chromosomes in Autism Spectrum Disorder. Am J Hum Genet. 2008;82:477–88. doi:10.1016/j.ajhg.2007.12.009.

8. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010;466:368–72. doi:10.1038/nature09146.

9. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, et al. Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. Neuron. 2011;70:863–85. doi:10.1016/j.neuron.2011.05.002.

10. Levy D, Ronemus M, Yamrom B, Lee Y ha, Leotta A, Kendall J, et al. Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders. Neuron. 2011;70:886–97. doi:10.1016/j.neuron.2011.05.015.

11. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet. 2014;94:677–94. doi:10.1016/j.ajhg.2014.03.018.

12. Youssoufian H, Pyeritz RE. Mechanisms and consequences of somatic mosaicism in humans. Nat Rev Genet. 2002;3:748–58. doi:10.1038/nrg906.

13. Freed D, Pevsner J. The Contribution of Mosaic Variants to Autism Spectrum Disorder.

PLoS Genet. 2016;12:e1006245. doi:10.1371/journal.pgen.1006245.
14. Dou Y, Yang X, Li Z, Wang S, Zhang Z, Ye AY, et al. Postzygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations. Hum Mutat. 2017;38:1002–13. doi:10.1002/humu.23255.
15. Lim ET, Uddin M, De Rubeis S, Chan Y, Kamumbu AS, Zhang X, et al. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. Nat Neurosci. 2017. doi:10.1038/nn.4598.
16. King DA, Jones WD, Crow YJ, Dominiczak AF, Foster NA, Gaunt TR, et al. Mosaic structural variation in children with developmental disorders. Hum Mol Genet. 2015;24:2733–45. doi:10.1093/hmg/ddv033.
17. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. Nat Rev Genet. 2013;14:307–20. doi:10.1038/nrg3424.
18. Fernández LC, Torres M, Real FX. Somatic mosaicism: on the road to cancer. Nat Rev Cancer. 2015;16:43–55. doi:10.1038/nrc.2015.1.
19. Campbell IM, Shaw CA, Stankiewicz P, Lupski JR. Somatic mosaicism: implications for disease and transmission genetics. Trends Genet. 2015;31:382–92. doi:10.1016/j.tig.2015.03.013.
20. Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat Genet. 2012;44:642–50. doi:10.1038/ng.2271.
21. Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, Rodriguez-Santiago B, et al. Detectable clonal mosaicism and its relationship to aging and cancer. Nat Genet. 2012;44:651–8. doi:10.1038/ng.2270.
22. Žilina O, Koltšina M, Raid R, Kurg A, Tõnisson N, Salumets A. Somatic mosaicism for copy-neutral loss of heterozygosity and DNA copy number variations in the human genome. BMC Genomics. 2015;16:703. doi:10.1186/s12864-015-1916-3.
23. Rodriguez-Santiago B, Malats N, Rothman N, Armengol L, Garcia-Closas M, Kogevinas M, et al. Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. Am J Hum Genet. 2010;87:129–38. doi:10.1016/j.ajhg.2010.06.002.
24. Jamuar SS, Lam A-TN, Kircher M, D'Gama AM, Wang J, Barry BJ, et al. Somatic Mutations in Cerebral Cortical Malformations. N Engl J Med. 2014;371:733–43. doi:10.1056/NEJMoa1314432.
25. Bedrosian TA, Linker S, Gage FH. Environment-driven somatic mosaicism in brain disorders. Genome Med. 2016;8:58. doi:10.1186/s13073-016-0317-9.
26. Shirley MD, Baugher JD, Stevens BL, Tang Z, Gerry N, Beiswanger CM, et al. Chromosomal variation in lymphoblastoid cell lines. Hum Mutat. 2012;33:1075–86.
27. González JR, Rodríguez-Santiago B, Cáceres A, Pique-Regi R, Rothman N, Chanock SJ, et al. A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. BMC Bioinformatics. 2011;12:166. doi:10.1186/1471-2105-12-166.
28. Baugher JD, Baugher BD, Shirley MD, Pevsner J. Sensitive and specific detection of mosaic chromosomal abnormalities using the Parent-of-Origin-based Detection (POD) method. BMC Genomics. 2013;14:367. doi:10.1186/1471-2164-14-367.
29. R Development Core Team RFFSC. Computational Many-Particle Physics. Vienna Austria R Foundation for Statistical Computing. 2008;739:ISBN 3-900051-07-0. doi:10.1007/978-3-540-74686-7.
30. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45. doi:10.1101/gr.092759.109.
31. Machiela MJ, Zhou W, Sampson JN, Dean MC, Jacobs KB, Black A, et al. Characterization of large structural genetic mosaicism in human autosomes. Am J Hum Genet. 2015;96:487–97. doi:10.1016/j.ajhg.2015.01.011.
32. Machiela MJ, Chanock SJ. The ageing genome, clonal mosaicism and chronic disease. Curr Opin Genet Dev. 2017;42:8–13. doi:10.1016/j.gde.2016.12.002.
33. Conlin LK, Thiel BD, Bonnemann CG, Medne L, Ernst LM, Zackai EH, et al. Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. Hum Mol Genet. 2010;19:1263–75.
34. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al.

Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron. 2015;87:1215–33. doi:10.1016/j.neuron.2015.09.016.
35. Yurov YB, Iourov IY, Vorsanova SG, Liehr T, Kolotii AD, Kutsev SI, et al. Aneuploidy and confined chromosomal mosaicism in the developing human brain. PLoS One. 2007;2:e558. doi:10.1371/journal.pone.0000558.
36. Bushman DM, Chun J. The genomically mosaic brain: Aneuploidy and more in neural diversity and disease. Semin Cell Dev Biol. 2013;24:357–69. doi:10.1016/j.semcdb.2013.02.003.
37. Rehen SK. Constitutional Aneuploidy in the Normal Human Brain. J Neurosci. 2005;25:2176–80. doi:10.1523/JNEUROSCI.4560-04.2005.
38. Ma R, Deng L, Xia Y, Wei X, Cao Y, Guo R, et al. A clear bias in parental origin of de novo pathogenic CNVs related to intellectual disability, developmental delay and multiple congenital anomalies. Sci Rep. 2017;7:44446. doi:10.1038/srep44446.

CHAPTER 2

Robust estimation of mosaic loss of chromosome Y with genotype-array-intensity data

Juan R González, **Marcos López-Sánchez**, Pedro Puig, Tonu Esko, Luis A Pérez-Jurado. 2017

Submitted

Robust estimation of mosaic loss of chromosome Y with genotype-array-intensity data

Juan R González[1,2,3,4], Marcos López-Sánchez[1,3], Pedro Puig[4], Tonu Esko[5], Luis A Pérez-Jurado[3]

Abstract: Mosaic chromosomal loss of Y events detectable in blood are one of the most common rearrangements in elderly man associated with all-cause mortality, haematological cancer, solid tumors, vascular complications, Alzheimer and all-risk mortality. To avoid limitations in the methods used so far, we propose a new method called MADloy that performs a robust estimation of mosaic loss of chromosome Y events using genotype array intensity data. The proposed method uses a non-inverse gaussian model to detect alterations of the Log R Ratio values in the male-specific Y region, and shows a better estimation of abnormal events, detecting losses and gains in chromosome Y with a higher sensitivity than previous methods.

Post-zygotic mutations of any type (small and mid-size sequence changes and rearrangements, chromosomal events, retrotransposition events, or epigenetic marks) lead to somatic mosaicism, the coexistence of cells of distinct genetic composition within an organism. Somatic mosaicism is the rule in any multicellular organism including human beings, and has been established as a cause of miscarriage, birth defects and developmental delay (Lu et al. 2008; Conlin et al. 2010). Detectable clonal mosacisism for chromosomal rearrangements in blood is more common than previously thought (Rodriguez-Santiago et al. 2010) with a prevalence that increases with age (Jacobs et al. 2012) and confers an increased risk for haematological cancer and some solid tumors (Jacobs et al., Laurie et al., Wright et al.), increased risk of vascular complications in patients with type 2 diabetes (Bonnefond et al, 2013) and higher overall mortality risk (Forsberg et al. 2014). These findings have underscored the importance of considering the role and time-dependent nature of somatic mutations in the etiology of complex diseases.

The most common somatic mutation identified in humans is the loss of chromosome Y (LOY) in male individuals. Mosaic LOY (mLOY) is associated with smoking and its frequency increases with age, being also associated with increased risk of Alzheimer's disease (AD) (Dumanski et al. 2016) and major cardiovascular defects (Haitjema et al. 2017). and higher overall mortality (Fosberg et al, 2014).

However, the mechanisms that regulate LOY, and its clinical relevance, have received little attention so far, in part due to the difficulties of the identification of LOY at high scale. Zhou et al (2016) provided the first example of a common susceptibility locus for genetic mosaicism, specifically for mLOY, that maps on TCL1A gene (Zhou et al. 2016). Wright et al. identified 19 genomic regions that are associated with mLOY at genome-wide significance level (Wright et al. 2017). In the same study, the authors by performing additional epigenome-wide methylation analyses in whole blood, found 36 differentially methylated sites associated with mLOY.

The existing scientific evidences demonstrate that SNP array intensity data enables a measure of LOY at population scale. Therefore, the joint analysis of current large number of GWAS with epigenomic and transcriptomic data will help in the elucidation of genes implicated in aneuploidy, genome instability and several complex diseases such as cancer or any other age-related. Scalable and robust statistical methods as well as efficient bioinformatic tools are crucial to infer LOY status from thousands of existing SNP arrays. These tools should contain not only proper methods to analyze LRR data, but also, parallelizable functions to deal with huge amount of samples. To this end, MADloy package (**See Online Methods and Supplementary Material** 1) has been created. The package includes all the features that are next described.

The calling of LOY status using genotype-array intensity has been performed by following the method described in Forsberg et al. (2014) with some modificiations. The authors proposed to analyze the log R ratio (LRR) values of SNPs probes in the male-specific region of chromosome Y (mLRR-Y) in the 56-Mb region between pseudoautosomal regions 1 and 2 (PAR1 and PAR2) on chromosome Y (chrY:2,694,521-59,034,049, hg19/GRCh37). Then, by using a very ad-hoc method, individuals are LOY-scored based on a threshold that is defined as the lower limit of the 99% confidence interval of the experimentally induced mLRR-Y variation (**See Online Methods**). The authors assume that LRR follow a symmetrical distribution. By definition, LRR is the ratio between two intensities. It is well known that assuming normality in these situations may be incorrect since data use to be skewed. Figure 1 a) and plotNIG figures for EGCUT, KIRC and BLCA datasets in **Supplementary File** 1 illustrate that the distribution of LRR cannot be considered symmetrical. Actually, the distribution of mLRR-Y is also characterized for having a heavy-tailed behavior. Therefore, we proposed to use a Normal Inverse Gaussian (NIG) distribution which properly fits the ratio of two random variables capturing this type of oddities (Barndorff-Nielsen 1997; Barndorff-Nielsen and Prause 2001). Another consideration is the X transposed region (XTR) that is shared between X and Y chromosome. This region is removed from the analysis because XTR can be affected by alterations in chromosome X redefining the mLRR-Y region (chrY 6,611,498-24,510,581, hg19/GRCh37). (Zhou et al. 2016). Forsberg et al. (2014) also consider that large values of mLRR-Y should be kept in the analysis since they are not real gains of chromosome Y (XYY). This assumption may not hold since we have observed that in some cases these samples can be real gains (e.g XYY) (**See Online Methods**).

In an attempt to avoid the limitations that may appear when performing calling using a threshold rule, some authors proposed to include in the models the mean intensity of mLRR-Y as a continuous variable (Wright et al. 2017). However, there are scientific evidence that this approach has the limitation of reducing the power in association studies (Yang, Wray, and Visscher 2010). In particular, this limitation has been shown in genomic association settings where copy number variant status is estimated using continuous intensities (LRR, qPCR, MLPA, ...) (Redon et al. 2006; González et al. 2009).

Another important limitation of the methods used so far to analyze LOY is that calling procedures do not use information

about B deviation (Rodriguez-Santiago, 2010). We have noticed that the calling of those individuals having LRR close to the threshold is not properly performed. This may happen because partial gains or deletions can affect only one of the chromosomal arms, some of these affecting also the pseudoautosomal regions and its allelic balance, and remain hidden by the LRR values of the rest of the chromosome. Therefore, in those cases, providing further information about the B deviation in the pseudoautosomal regions can really improve the calling procedure **(see Supplementary File 3)**.

To examine the power of our proposed method, we run some simulation studies where our proposed method is compared with the method proposed by Forsberg et al (2014) and the quantitative approach proposed by Wright et al. (2017) (**See Online Methods**). Figure 2 illustrates how analyzing LOY status as dichotomous variable (normal vs LOY) outperforms the analysis of considering mLRR-Y as a continuous variable. Additionally, the figure also demonstrates that the threshold method is also having less power than the method based on using NIG to estimate LRR background noise distribution. These results correspond to the association analysis of LOY and quantitative traits (i.e. age). Simulation results of association analyses between quantitative traits (e.g. case/control) and LOY are shown in the different **Figures of Supplementary File 2** and the same conclusions can be achieved.

The practical usefulness of MADloy pipeline and our proposed method to calling LOY is illustrated by analyzing hundreds of samples from different settings. We analyze samples from general population (EGCUT biobank), cancer data (KIRC and BLCA tumors from TCGA project) and individuals diagnosed with Alzheimer's disease (NIA cohort) (**see Online Methods**). By handling these three different types of data we cover different situations that can appear when analyzing LOY. That is, we cover: 1) data from general population or specific complex diseases; 2) data having different degree of homogeneity; 3) data being measured in different tissues; and 4) data obtained from different platforms.

One of the main findings published in the literature is that LOY is associated with age. The analysis of EGCUT data further provides evidences on this finding. When analyzing mLRR-Y data as dichotomous variable (normal/LOY) using our proposed calling method, we observe that LOY appear in older samples ($p = 7.9x10^{-4}$) as described by Dumanski et al. (2106). However, the analysis using the threshold method ($p = 0.0156$) and mLRR as a quantitative variable ($p = 1.2x10^{-3}$) does not provide so strong evidence (**Section 2.2 in Supplementary File 1**). Similar analysis performed in TCGA data showed some discrepancies depending how LOY information is included in the models. KIRC dataset also provides evidence of association between LOY and age ($p=0.0002$ and $p=0.0003$ using categorical or continuous assessment of LOY). However, in the LGG dataset, the effect of age is only observed when LOY is analyzed as categorical variable (normal/LOY) ($p=0.0089$) (**Section 3.4 in Supplementary File 1**). All these results are in the same direction as our simulation studies. .

The association analysis between LOY and cancer using TCGA data also reveals that the analysis of LOY should be performed by considering a normal/LOY variable instead of using continuous mLRR measurement as a surrogate. The association between LOY

and KIRC showed a highly significant association between LOY and cancer samples. Although all methodologies are providing very low p-values since there is a strong correlation between cancer and LOY, again, our proposed method is providing the strongest evidence (p-value=$1.1x10^{-77}$) compared to threshold method (p-value=$1.3x10^{-17}$) and the quantitative approach (p-value=$1.6x10^{-14}$). The analysis of LGG dataset is also providing similar conclusions since the p-value obtained from quantitative assessment of LOY is having a significant p-value with lower order of magnitude than the methods based on considering LOY as a binary variable (**Section 3.3 in Supplementary File 1**).

There is also evidence of association between AD and LOY. Dumanksi et al (2017) showed that men with LOY at blood sampling had an elevated risk of incident AD diagnosis during follow-up time (hazard ratio [HR] = 6.80, p=0.0011). The analysis of 644 samples belonging to NIA cohort reveals that the risk of developing AD is twice in LOY carriers (HR=1.96, p=0.0044, AD events=278) (**Section 4 in Supplementary File 1**).

To assess the validity of the obtained results, 58 samples (exome data) from TCGA dataset were downloaded in order to estimate its copy-number state in the mLRR-Y as proposed by (Forsberg et al. 2014) by using FREEC-control software. The results obtained in the validation of those samples showed a coincidence classification ratio of 93,1% (N=58, concordant=54). Of these, the number of misclassifications affects mainly those events classified with exome as "normal" (N=3), but classified with our method as "LOY" (N=1) or "XYY" (N=2). The remaining misclassification is a "XYY" event in exome classified as "normal" with our proposed method (**Supplementary Table** 1).

An indirect validation of our proposed methodology can be performed by analyzing transcriptomic data. It is expected that cases having LOY decrease gene expression in Y chromosome. The analysis of KIRC and BLCA datasets with regard to RNAseq data showed that the top-10 downregulated genes associated with LOY include **TTTY**4C, UTY TMSB4Y, USP9Y, ZFY, EIF1AY, RSP4Y1, **TTTY**15 **(Section 5 in Supplementary File 1)**, agreeing with the expected differences when comparing individuals with a chromosome Y (male) against individuals without a chromosome Y (female).

As a conclusion, we have provided a number of evidences that genotype-array-intensity data can be used in LOY association studies where different outcomes can be considered. We have also shown that using NIG to fit the background distribution of LRR data increases the power of detecting positive associations compared to existing methodologies. The method and the bioinformatic tool presented in this work also illustrate that the proposed workflow is simple, interpretable, and fast. Having a method with these characteristics is essential to anticipate the avenue of large number of individuals and to characterize those individuals with high sensitivity and specificity, enabling to decipher the mechanism or study the role of LOY in complex diseases MADloy will allow researchers to re-analyze thousands of existing GWAS data in public available repositories and integrate LOY data with transcriptome, epigenome or any other omic data that are currently being generated in large epidemiological studies.

References

Barndorff-Nielsen, Ole E. 1997. "Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling." *Scandinavian Journal of Statistics* 24 (1): 1–13. doi:10.1111/1467-9469.00045.

Barndorff-Nielsen, Ole E., and Karsten Prause. 2001. "Apparent Scaling." *Finance and Stochastics* 5 (1): 103–13. doi:10.1007/s007800000020.

Conlin, Laura K., Brian D. Thiel, Carsten G. Bonnemann, Livija Medne, Linda M. Ernst, Elaine H. Zackai, Matthew A. Deardorff, Ian D. Krantz, Hakon Hakonarson, and Nancy B. Spinner. 2010. "Mechanisms of Mosaicism, Chimerism and Uniparental Disomy Identified by Single Nucleotide Polymorphism Array Analysis." *Human Molecular Genetics* 19 (7): 1263–75. doi:10.1093/hmg/ddq003.

Dumanski, Jan P., Jean Charles Lambert, Chiara Rasi, Vilmantas Giedraitis, Hanna Davies, Benjamin Grenier-Boley, Cecilia M. Lindgren, et al. 2016. "Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease." *American Journal of Human Genetics* 98 (6): 1208–19. doi:10.1016/j.ajhg.2016.05.014.

Forsberg, Lars A, Chiara Rasi, Niklas Malmqvist, Hanna Davies, Saichand Pasupulati, Geeta Pakalapati, Johanna Sandgren, et al. 2014. "Mosaic Loss of Chromosome Y in Peripheral Blood Is Associated with Shorter Survival and Higher Risk of Cancer." *Nature Genetics* 46 (6): 624–28. doi:10.1038/ng.2966.

González, Juan R, Isaac Subirana, Geòrgia Escaramís, Solymar Peraza, Alejandro Cáceres, Xavier Estivill, and Lluís Armengol. 2009. "Accounting for Uncertainty When Assessing Association between Copy Number and Disease: A Latent Class Model." *BMC Bioinformatics* 10 (1): 172. doi:10.1186/1471-2105-10-172.

Haitjema, Saskia, Daniel Kofink, Jessica van Setten, Sander W. van der Laan, Arjan H. Schoneveld, James Eales, Maciej Tomaszewski, et al. 2017. "Loss of Y Chromosome in Blood Is Associated With Major Cardiovascular Events During Follow-Up in Men After Carotid EndarterectomyCLINICAL PERSPECTIVE." *Circulation: Cardiovascular Genetics* 10 (4): e001544. doi:10.1161/CIRCGENETICS.116.001544.

Jacobs, Kevin B, Meredith Yeager, Weiyin Zhou, Sholom Wacholder, Zhaoming Wang, Benjamin Rodriguez-Santiago, Amy Hutchinson, et al. 2012. "Detectable Clonal Mosaicism and Its Relationship to Aging and Cancer." *Nature Genetics* 44 (6). Nature Publishing Group: 651–58. doi:10.1038/ng.2270.

Lu, X.-Y., M. T. Phung, C. A. Shaw, K. Pham, S. E. Neil, A. Patel, T. Sahoo, et al. 2008. "Genomic Imbalances in Neonates With Birth Defects: High Detection Rates by Using Chromosomal Microarray Analysis." *Pediatrics* 122 (6): 1310–18. doi:10.1542/peds.2008-0297.

Redon, Richard, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, et al. 2006. "Global Variation in Copy Number in the Human Genome." *Nature* 444 (7118): 444–54. doi:10.1038/nature05329.

Rodríguez-Santiago, Benjamín, Núria Malats, Nathaniel Rothman, Lluís Armengol, Montse Garcia-Closas, Manolis Kogevinas, Olaya Villa, et al. 2010. "Mosaic Uniparental Disomies and Aneuploidies as Large Structural Variants of the Human Genome."

American Journal of Human Genetics 87 (1). Elsevier: 129–38. doi:10.1016/j.ajhg.2010.06.002.

Ross, M T, D V Grafham, A J Coffey, S Scherer, K McLay, D Muzny, M Platzer, et al. 2005. "The DNA Sequence of the Human X Chromosome." *Nature* 434 (March): 325–37. doi:nature03440 [pii]\r10.1038/nature03440.

Wright, Daniel J, Felix R Day, Nicola D Kerrison, Florian Zink, Alexia Cardona, Patrick Sulem, Deborah J Thompson, et al. 2017. "Genetic Variants Associated with Mosaic Y Chromosome Loss Highlight Cell Cycle Genes and Overlap with Cancer Susceptibility." *Nature Genetics*, no. March. Nature Publishing Group: 1–8. doi:10.1038/ng.3821.

Yang, Jian, Naomi R. Wray, and Peter M. Visscher. 2010. "Comparing Apples and Oranges: Equating the Power of Case-Control and Quantitative Trait Association Studies." *Genetic Epidemiology* 34 (3): 254–57. doi:10.1002/gepi.20456.

Zhou, Weiyin, Mitchell J Machiela, Neal D Freedman, Nathaniel Rothman, Nuria Malats, Casey Dagnall, Neil Caporaso, et al. 2016. "Mosaic Loss of Chromosome Y Is Associated with Common Variation near TCL1A." *Nature Genetics* 48 (5): 563–68. doi:10.1038/ng.3545.

CHAPTER 3

Nested Inversions Polymorphisms predispose chromosome 22q11.2 to meiotic rearrangements

Demaerel W, Hestand MS, Vergaelen E, Swillen A, López-Sánchez M, Pérez-Jurado LA, et al. RETRACTED: Nested Inversion Polymorphisms Predispose Chromosome 22q11.2 to Meiotic Rearrangements. Am J Hum Genet. 2017 Oct 5;101(4):616–22. DOI: 10.1016/j.ajhg.2017.09.002

DISCUSSION

Chromosomal mosaic events and ancestral polymorphic inversions, two different types of genetics variants, contribute significantly to autism spectrum disorders (ASD).

The results obtained in the analysis of chromosomal mosaicism, available in chapter 1, are that a small autosomal chromosomal mosaicism rate was detected in ASD probands blood samples (0.45%). The majority of these events (80%, 16/20) were undetected in previous copy-number analyses of the same data (Pinto et al. 2010; S. J. Sanders et al. 2011), and are likely pathogenic. Regarding the contribution of nine analysed ancestral polymorphic inversions, only two, inv8p23.1 and inv17q21.31 have a statistical significant association with ASD (inv17q21.31, p = 1.14E-02; inv8p23.1, p = 9.34E-03), and an over-transmission compared to healthy siblings. The association of these two ancestral polymorphic inversions contributes also to explain part of the genetic common variation to ASD.

The fact that both genetic variants are associated to autism spectrum disorders supports the *neuroconstructivism* theory in neurodevelopmental disorders. In one hand chromosomal mosaic events cause an alteration of a genomic region in an early developmental stage, affecting brain development. In the other hand, ancestral polymorphic inversions alleles modulate gene expression of the genes contained in its region since the first

division. This theory also agrees with the fact that no specific brain or functionality is exclusively affected in Autism spectrum disorders, and with the shared comorbidity and genetic factors with other neurodevelopmental disorders (Mitchell 2011; Bralten et al. 2017)

Autism spectrum disorder has been reported as a complex and heterogeneous disorder (Hu, Chahrour, and Walsh 2014; Mitchell 2011), and the fact that the chromosomal mosaic events detected in ASD patients were not homogeneous are in agreement, with ASD genetic complexity. In addition, the consequences of chromosomal mosaic variants are unknown in brain development, but inversions, and specifically inv17q21.31, were reported to affect gene expression (de Jong et al. 2012) that goes in the same direction as the differential expression found in *MAPT* and *CRHR1* for frontal cortex and cerebellum tissues from GTEX analysis (Gutiérrez-Arumí 2016). In addition to these changes, the effect on physiological aneuploidy and L1 mediated rearrangements dynamics that takes place in brain (Pack et al. 2005; Yurov et al. 2007) are still subject of debate (Andriani, Vijg, and Montagna 2017).

When considering the heritability of autism in monozygotic twins, there is still debate about its true estimation, but the conservative approach is 50% (Bourgeron 2016). This heritability includes ancestral polymorphic variants but not chromosomal mosaicism, because chromosomal mosaic events are not inherited and could be

accounted to environmental factors, as well as somatic point mutations. However, gonadal mosaicism in parents can trigger rescue mechanisms (trisomic rescue, compensatory for UPD, nondisjunction, premature separation of sister chromatids, reverse segregation, nondisjunction or anaphase lagging for aneuploidies), that triggers the arise of a mosaic population (Sandin et al. 2014).

Focusing on the endophenotypes of autism, in the analysis of chromosomal mosaic events we expected a higher chromosomal mosaic rate in simplex or idiopathic cases (0.44%, 14/3152) versus multiplex or familiar (0.86%, 3/347), but the rates are not statistically different due to the low number of multiplex cases. The study of endophenotypes in the two over-transmitted inversions inv8p23 and inv17q21.31 showed a stronger association with high functionality and multiplex cases (Gutiérrez-Arumí 2016). The association with multiplex cases agrees with fact that inversions are common inherited variants, and is consistent with the polygenic risk model (Gaugler et al. 2014).

One of the characteristics of ASD is a discussed 4:1 male to female ratio (Loomes, Hull, and Mandy 2017; Crow 2000), where the lower female ratio has been hypothesized to be caused by a protective female effect (Chaste, Roeder, and Devlin 2017). The chromosomal mosaic rates differences between gender are not statistically different (females - 0.33%, 2/595; males - 0.42%, 16/3810) which does not agree with the increased prevalence, suggesting that chromosomal mosaic events are not affected by the

female protective effect hypothesis. When considering the parental origin of the event for gains and losses, the results obtained are in agreement with the results obtained in the analysis of copy-number variants by Sanders (2015) (S. J. Sanders et al. 2015), where gains are originated (100% gains with maternal origin).

In the light of the chromosomal mosaicism rates in general population (Machiela et al. 2015; Jacobs et al. 2012) for the same age bin as ASD probands analysed, chromosomal mosaicism stands out as a very rare event. The results presented in this work demonstrated that the low incidence of these genetic variants in ASD is higher than in unaffected siblings of the same age. However, one of the limitations of previous studies was the chromosomal mosaic event sizes, limited by the SNP arrays used. The availability of high density SNP array data, with more than 1 million of positions analysed, allows a theoretical effective resolution of 6 Kb (Illumina 2010), but in order to have high-confidence results, we considered a threshold resolution of 400 Kb.

One of the most controversial topics is the relationship between chromosomal mosaicism events detected in blood and its presence in other tissues, and specifically for this work, in brain. The results exposed does not show experimental validation in brain tissue or other tissues with ectodermal origin. Even so, previous reports on developmental disorders and mosaicism (D. A. King et al. 2015), joint with the age bin of the probands (4-18 years old) and the ratio of cells affected suggest an embryonic origin, and not a

chromosomal instability, carcinogenic event or myelodysplastic syndrome, where no concordance has been reported (Blatt, Deal, and Mesibov 2010). even when some ASD mutations have been related to oncogenic processes (Darbro et al. 2016).

Regarding the mechanisms implicated in the arise of chromosomal events, there is only one autosomal event that affects a whole chromosome, which is a uniparental isodisomy. The most probable mechanism involved is a trisomic rescue or compensatory UPD, with parental origin (W. P. Robinson 2000). The other uniparental isodisomic events detected could have originated via somatic recombination, break-induced replication or nondisjunction, but it cannot be clearly determined. Of the remaining events, all the autosomal gains are explained by maker chromosomes generation, while autosomal losses 6 out of 7 are interstitial and could have been generated by break-induced replication or alternative nonhomologous end joining mechanisms.

Due to the fact that the mosaic events detected have arisen postzygotically, and therefore not present in parents and *de novo*, the number of events that can have consequences in parental transmission is only limited to aneuploid gametes or mutation that affects chromosomal stability which agrees with the link between chromosomal instability and aneuploidy (Thompson and Compton 2008). This chromosomal instability has been reported as a common postzygotic event in cleavage-stage embryos (Vanneste et al. 2009) and it explains the low human fecundity. In line with these facts,

preimplantation genetic screening for embryonic mosaicism has become a routine analysis in preimplantation embryos (Capalbo and Rienzi 2017).

About the detection of mosaic loss of chromosome Y (mLOY), a new bioinformatic method been developed that replicate previous results and improves their methods (Forsberg et al. 2014; Jan P. Dumanski et al. 2016; Haitjema et al. 2017) in two essential points. The first point is analysing the male-specific region of chromosome Y without the X transposed region. The X transposed region is 99% identical to Xq21 region, and thus cannot be considered as male-specific due to its copy-number state of 2 (Cotter, Brotman, and Wilson Sayres 2016). The second point is the use a robust model to detect alterations in the male-specific Y Log R Ratio (LRR) by taking into account the LRR value distribution and the median Log R Ratio of this region (mLRR-Y). The previous methods consider a normal distribution of these values, which has been proven to better adjust with a non-inverse Gaussian distribution.

In addition to the methodological approach proposed, the experimental detection of chromosomal mosaic events in ASD, with the copy-number estimation obtained by the B Allele Frequency (BAF) values, allowed to establish a mathematical relationship between LRR and copy-number estate for Illumina arrays. This relationship is one of the tools that enables the transformation of log r ratio values from a copy-number state of 2 to a copy-number state of 1 or vice-versa. This transformation is required to transform LRR

values of Y chromosome computed with a default ploidy of 2 to a default ploidy of 1 in those cases where the Illumina GenomeStudio software have not done it.

When performing association analyses, results show that the use of a categorical classification of the mLOY events (normal, LOY, GOY) outperforms the use of continuous values of mLRR-Y in terms of statistical power, which agrees with the arguments in Yang et al (2010) (Yang, Wray, and Visscher 2010). The addition of an advanced analysis of the results where the B Allele frequency is measured in the pseudoautosomal regions of chromosome X and Y adds a higher degree of confidence in the mLOY calls, a procedure that was also performed in the literature methods. This method can be used to study the importance of Y chromosome loss in sex-biased disorders like autoimmune diseases where mLOY is described (Persani et al. 2012), Parkinson disease (Forsberg 2017) and haematological disorders as example.

Regarding the effect and study of ancestral polymorphic inversions, while the two inversions found associated with ASD (inv8p23.1, inv17q21.31) have been extensively studied and methods are well established, the study of putative inversions detected by *inveRsion* software show a higher degree of difficulty, especially in inversions where clusters suggest more than two alleles (inv2q13, inv4q13..2, inv15q24.2, inv16p11.2), meaning that probably inv8p23.1 and inv17q21.31 are the exception rather than the rule. In addition, experimental validation in some of the putative inversions by

Interphase FISH have not showed conclusive results, except in the study of inv22q11.21. This inversion seems to play a mediator role in the microdeletion and microduplication syndrome on the 22q11.21 region, leading to DiGeorge/Velocardiofacial syndrome (ANNEX 2) (Demaerel et al. 2017), similar to the inversion in 7q11.23 region that mediates the microdeletion of region 7q11.23 in 1/3 of the Williams-Beuren syndrome cases (Osborne et al. 2001). The use of single cell strand-seq methods to detect chromosomal inversions has led to identify 111 inversions some of which were already known (A. D. Sanders et al. 2016), and sequencing methods that do not relies in identifying the strand fails to capture inversion events between large segmental duplications, even long reads (Eslami Rasekh et al. 2017).

The results obtained of the inv8p23.1 and inv17q21.31 show an over-transmission and association of the inverted alleles in ASD (Gutiérrez-Arumí 2016). The fact that association analyses, transmission studies in trios, and differential expression of genes for inversions genotype agrees with the message that these inversions plays an important role in the common susceptibility to ASD. These results are also in consonance with the association of inv8p23.1 with neuroticism and personality traits (Lo et al. 2017).

One of the main interests in the study of inv22q11.21 is the fact that DiGerorge/Velocardiofacial Syndrome is associated with autism and schizophrenia, which are determined by the variation present in the *PRODH* and *COMT* genes included in the region. The results

obtained in the analysis of the putative inversion by Fiber-FISH, identified two inversion alleles, one located between segmental duplication blocks of LCR22C and LCR22D, and another between blocks LRC22B and LRC22D (Figure 12). Something to take into consideration is that when studying the number of inverted fibers and non-inverted fibers by Fiber-FISH, no parent was found homozygous for the inversion, and none of the seven probands with the most common deletion between blocks LRC22A and LRC22D have any inverted fiber. It was discussed that the presence of the inversion disrupts the function of essential genes, which agrees with the gene disruption theory (Puig, Castellano, et al. 2015). However computational predictions of inversion allele frequencies, and the selective pressure will have reduced the presence of the inverted allele in population, disagrees with this hypothesis.

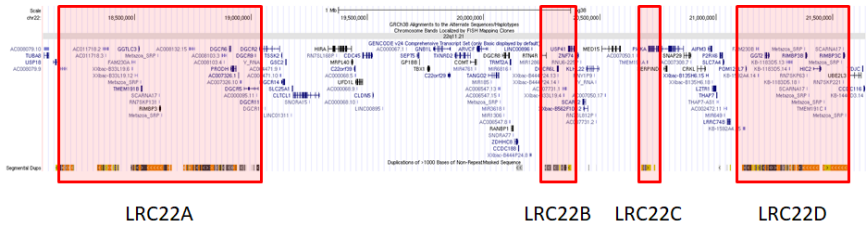


Figure 12: Structure of the 22q11.21 region in the reference genome

This figure describes the 22q11.21 region linked with DiGeorge Syndrome (hg19) obtained from UCSC genome viewer. The red boxes indicate the segmental duplication blocks in the region.

Taking into account the number of inversions detected by single cell strand-seq techniques, a large fraction of the detected inversions probably are modern inversions that have not accumulated enough sequence divergence to be characterized using the SNP techniques

used in this work. In addition, the paper of segmental duplications in the generation of new inversions seems of utmost importance (Dennis et al. 2017), and there is a lot of work in this field to do in order to identify the mechanisms of fixation of an inversion allele, as well as the current inversion load that are shaping the evolution of our species.

CONCLUSIONS

In the present dissertation chromosomal mosaic rearrangements and ancestral polymorphic inversions are analysed in the context of neurodevelopmental disorders:

- Chromosomal mosaic events detectable in blood samples are responsible for a small but significant proportion of patients with ASD (0.45%), implying that the contribution of somatic mutation to ASD is more important than previously expected.
- Detection of chromosomal mosaic events contributes novel genes and genomic regions that are involved in ASD aetiology.
- The use of two methods to detect chromosomal mosaic events (MAD and triPOD) allowed a better interpretation of the SNP array data and their processing. This understanding lead to the improvement of the MAD algorithm and the creation of a mosaic event simulator.
- We have optimized bioinformatic methods and algorithms to detect and quantify Loss of chromosome Y using data from SNP arrays, generating MADloy as a tool available for the scientific community (Through github at https://github.com/isglobal-brge/MADloy).
- MADloy over-performs in comparison to existing methods, and can be used to analyse large datasets in a fast manner. A derived method recently developed by our group (not

included in this thesis) is able to detect loss of chromosome Y using NGS data.

- Two ancestral polymorphic inversions, inv8p23.1 and in17q21.31, are associated with autism risk, in agreement with previous results showing over-transmission of risk alleles to autistic probands from parents.
- Improvements of the invClust method to genotype ancestral inversions using SNP array data allowed the definition of multiallelic inversions to better define allelic association studies in neurodevelopmental and other disorders.
- Computational prediction revealed a novel multiallelic inversion in 22q11.21 region which has then been validated by fiber-FISH. Heterozygosis for this inversion is a predisposing factor for the generation of the most common recurrent microduplications and microdeletions, with phenotypic implications in neurodevelopmental disorders.

REFERENCES

Abyzov, Alexej, Livia Tomasini, Bo Zhou, Nikolaos Vasmatzis, Gianfilippo Coppola, Mariangela Amenduni, Reenal Pattni, et al. 2017. "One Thousand Somatic SNVs per Skin Fibroblast Cell Set Baseline of Mosaic Mutational Load with Patterns That Suggest Proliferative Origin." *Genome Research* 27 (4). Cold Spring Harbor Laboratory Press: 512–23. doi:10.1101/gr.215517.116.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*. *Arlington*. doi:10.1176/appi.books.9780890425596.744053.

Andriani, Grasiella A., Jan Vijg, and Cristina Montagna. 2017. "Mechanisms and Consequences of Aneuploidy and Chromosome Instability in the Aging Brain." *Mechanisms of Ageing and Development* 161: 19–36. doi:10.1016/j.mad.2016.03.007.

Antonacci, Francesca, Jeffrey M Kidd, Tomas Marques-Bonet, Mario Ventura, Priscillia Siswara, Zhaoshi Jiang, and Evan E Eichler. 2009. "Characterization of Six Human Disease-Associated Inversion Polymorphisms." *Human Molecular Genetics* 18 (14). Oxford University Press: 2555–66. doi:10.1093/hmg/ddp187.

Azcona, C, P Bareille, and R Stanhope. 1999. "Lesson of the Week: Turner's Syndrome Mosaicism in Patients with a Normal Blood Lymphocyte Karyotype." *BMJ (Clinical Research Ed.)* 318 (7187). BMJ Publishing Group: 856–57. http://www.ncbi.nlm.nih.gov/pubmed/10092267.

Baugher, Joseph D, Benjamin D Baugher, Matthew D Shirley, and Jonathan Pevsner. 2013. "Sensitive and Specific Detection of Mosaic Chromosomal Abnormalities Using the Parent-of-Origin-Based Detection (POD) Method." *BMC Genomics* 14 (1): 367. doi:10.1186/1471-2164-14-367.

Biesecker, Leslie G., and Nancy B. Spinner. 2013. "A Genomic View of Mosaicism and Human Disease." *Nature Reviews. Genetics* 14 (5). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 307–20.

doi:10.1038/nrg3424.

Blatt, Julie, Allison M. Deal, and Gary Mesibov. 2010. “Autism in Children and Adolescents with Cancer.” *Pediatric Blood & Cancer* 54 (1): 144–47. doi:10.1002/pbc.22303.

Bourgeron, Thomas. 2016. “Current Knowledge on the Genetics of Autism and Propositions for Future Research.” *Comptes Rendus Biologies* 339 (7–8): 300–307. doi:10.1016/j.crvi.2016.05.004.

Bralten, J, K J van Hulzen, M B Martens, T E Galesloot, A Arias Vasquez, L A Kiemeney, J K Buitelaar, J W Muntjewerff, B Franke, and G Poelmans. 2017. “Autism Spectrum Disorders and Autistic Traits Share Genetics and Biology.” *Molecular Psychiatry*, May. doi:10.1038/mp.2017.98.

Brouha, B., J. Schustak, R. M. Badge, S. Lutz-Prigge, A. H. Farley, J. V. Moran, and H. H. Kazazian. 2003. “Hot L1s Account for the Bulk of Retrotransposition in the Human Population.” *Proceedings of the National Academy of Sciences* 100 (9): 5280–85. doi:10.1073/pnas.0831042100.

Burrell, Rebecca A., Nicholas McGranahan, Jiri Bartek, and Charles Swanton. 2013. “The Causes and Consequences of Genetic Heterogeneity in Cancer Evolution.” *Nature* 501 (7467): 338–45. doi:10.1038/nature12625.

Cáceres, Alejandro, and Juan R. González. 2015. “Following the Footprints of Polymorphic Inversions on SNP Data: From Detection to Association Tests.” *Nucleic Acids Research* 43 (8): e53. doi:10.1093/nar/gkv073.

Cáceres, Alejandro, Suzanne S Sindi, Benjamin J Raphael, Mario Cáceres, and Juan R González. 2012. “Identification of Polymorphic Inversions from Genotypes.” *BMC Bioinformatics* 13 (1): 28. doi:10.1186/1471-2105-13-28.

Campbell, Ian M., Chad A. Shaw, Pawel Stankiewicz, and James R. Lupski. 2015. “Somatic Mosaicism: Implications for Disease and Transmission Genetics.” *Trends in Genetics* 31 (7): 382–92. doi:10.1016/j.tig.2015.03.013.

Capalbo, Antonio, and Laura Rienzi. 2017. “Mosaicism between Trophectoderm and Inner Cell Mass.” *Fertility and Sterility* 107 (5): 1098–1106. doi:10.1016/j.fertnstert.2017.03.023.

Chaste, Pauline, Kathryn Roeder, and Bernie Devlin. 2017. “The Yin and Yang of Autism Genetics: How Rare De Novo and Common Variations Affect Liability.” *Annual Review of Genomics and Human Genetics* 18 (1): annurev-genom-083115-022647. doi:10.1146/annurev-genom-083115-022647.

Chatterjee, Nimrat, and Graham C. Walker. 2017. “Mechanisms of DNA Damage, Repair, and Mutagenesis.” *Environmental and Molecular Mutagenesis* 58 (5): 235–63. doi:10.1002/em.22087.

Cheng, Jiqiu, Evelyne Vanneste, Peter Konings, Thierry Voet, Joris R Vermeesch, and Yves Moreau. 2011. “Single-Cell Copy Number Variation Detection.” *Genome Biology* 12 (8). BioMed Central: R80. doi:10.1186/gb-2011-12-8-r80.

Conover, Hailey N., and Juan Lucas Argueso. 2016. “Contrasting Mechanisms of de Novo Copy Number Mutagenesis Suggest the Existence of Different Classes of Environmental Copy Number Mutagens.” *Environmental and Molecular Mutagenesis* 57 (1): 3–9. doi:10.1002/em.21967.

Cotter, Daniel J., Sarah M. Brotman, and Melissa A. Wilson Sayres. 2016. “Genetic Diversity on the Human X Chromosome Does Not Support a Strict Pseudoautosomal Boundary.” *Genetics* 203 (1): 485–92. doi:10.1534/genetics.114.172692.

Crow, J F. 2000. “The Origins, Patterns and Implications of Human Spontaneous Mutation.” *Nature Reviews. Genetics* 1 (1): 40–47. doi:10.1038/35049558.

D’Souza, Hana, and Annette Karmiloff-Smith. 2017. “Neurodevelopmental Disorders.” *Wiley Interdisciplinary Reviews: Cognitive Science* 8 (1–2): e1398. doi:10.1002/wcs.1398.

Darbro, Benjamin W., Rohini Singh, M. Bridget Zimmerman, Vinit B. Mahajan, and Alexander G. Bassuk. 2016. “Autism Linked to Increased Oncogene Mutations but Decreased Cancer Rate.” *PLoS ONE* 11 (3). doi:10.1371/journal.pone.0149041.

de Jong, Simone, Iouri Chepelev, Esther Janson, Eric Strengman, Leonard H van den Berg, Jan H Veldink, and Roel A Ophoff. 2012. “Common Inversion Polymorphism at 17q21.31 Affects Expression of Multiple Genes in Tissue-Specific Manner.”

BMC Genomics 13 (1): 458. doi:10.1186/1471-2164-13-458.

de la Torre-Ubieta, Luis, Hyejung Won, Jason L Stein, and Daniel H Geschwind. 2016. "Advancing the Understanding of Autism Disease Mechanisms through Genetics." *Nature Medicine* 22 (4): 345–61. doi:10.1038/nm.4071.

Delhanty, J.D.A. 2011. "Inherited Aneuploidy: Germline Mosaicism." *Cytogenetic and Genome Research* 133 (2–4): 136–40. doi:10.1159/000323606.

Demaerel, Wolfram, Matthew S Hestand, Elfi Vergaelen, Ann Swillen, Marcos López-Sánchez, Luis A Pérez-Jurado, Donna M McDonald-McGinn, et al. 2017. "Nested Inversion Polymorphisms Predispose Chromosome 22q11.2 to Meiotic Rearrangements." *American Journal of Human Genetics* 101 (4). Elsevier: 616–22. doi:10.1016/j.ajhg.2017.09.002.

Dennis, Megan Y., Lana Harshman, Bradley J. Nelson, Osnat Penn, Stuart Cantsilieris, John Huddleston, Francesca Antonacci, et al. 2017. "The Evolution and Population Diversity of Human-Specific Segmental Duplications." *Nature Ecology & Evolution* 1 (3): 69. doi:10.1038/s41559-016-0069.

Di Noia, Javier M., and Michael S. Neuberger. 2007. "Molecular Mechanisms of Antibody Somatic Hypermutation." *Annual Review of Biochemistry* 76 (1): 1–22. doi:10.1146/annurev.biochem.76.061705.090740.

Dong, Shan, Michael F. Walker, Nicholas J. Carriero, Michael DiCola, A. Jeremy Willsey, Adam Y. Ye, Zainulabedin Waqar, et al. 2014. "De Novo Insertions and Deletions of Predominantly Paternal Origin Are Associated with Autism Spectrum Disorder." *Cell Reports* 9 (1): 16–23. doi:10.1016/j.celrep.2014.08.068.

Dou, Yanmei, Xiaoxu Yang, Ziyi Li, Sheng Wang, Zheng Zhang, Adam Yongxin Ye, Linlin Yan, et al. 2017. "Postzygotic Single-Nucleotide Mosaicisms Contribute to the Etiology of Autism Spectrum Disorder and Autistic Traits and the Origin of Mutations." *Human Mutation* 38 (8): 1002–13. doi:10.1002/humu.23255.

Dumanski, J. P., C. Rasi, M. Lonn, H. Davies, M. Ingelsson, V. Giedraitis, L. Lannfelt, et al. 2015. "Smoking Is Associated

with Mosaic Loss of Chromosome Y." *Science* 347 (6217): 81–83. doi:10.1126/science.1262092.

Dumanski, Jan P., Jean Charles Lambert, Chiara Rasi, Vilmantas Giedraitis, Hanna Davies, Benjamin Grenier-Boley, Cecilia M. Lindgren, et al. 2016. "Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease." *American Journal of Human Genetics* 98 (6): 1208–19. doi:10.1016/j.ajhg.2016.05.014.

Duncan, Andrew W. 2013. "Aneuploidy, Polyploidy and Ploidy Reversal in the Liver." *Seminars in Cell and Developmental Biology*. doi:10.1016/j.semcdb.2013.01.003.

Ehninger, Dan, Weidong Li, Kevin Fox, Michael P. Stryker, and Alcino J. Silva. 2008. "Reversing Neurodevelopmental Disorders in Adults." *Neuron*. doi:10.1016/j.neuron.2008.12.007.

Elsabbagh, Mayada, Gauri Divan, Yun Joo Koh, Young Shin Kim, Shuaib Kauchali, Carlos Marcín, Cecilia Montiel-Nava, et al. 2012. "Global Prevalence of Autism and Other Pervasive Developmental Disorders." *Autism Research* 5 (3): 160–79. doi:10.1002/aur.239.

Engle, Elizabeth C. 2010. "Human Genetic Disorders of Axon Guidance." *Cold Spring Harbor Perspectives in Biology*. doi:10.1101/cshperspect.a001784.

Eslami Rasekh, Marzieh, Giorgia Chiatante, Mattia Miroballo, Joyce Tang, Mario Ventura, Chris T. Amemiya, Evan E. Eichler, Francesca Antonacci, and Can Alkan. 2017. "Discovery of Large Genomic Inversions Using Long Range Information." *BMC Genomics* 18 (1): 65. doi:10.1186/s12864-016-3444-1.

Fernández, Luis C., Miguel Torres, and Francisco X. Real. 2015. "Somatic Mosaicism: On the Road to Cancer." *Nature Reviews Cancer* 16 (1). Nature Publishing Group: 43–55. doi:10.1038/nrc.2015.1.

Forsberg, Lars A. 2017. "Loss of Chromosome Y (LOY) in Blood Cells Is Associated with Increased Risk for Disease and Mortality in Aging Men." *Human Genetics*. doi:10.1007/s00439-017-1799-2.

Forsberg, Lars A., David Gisselsson, and Jan P. Dumanski. 2016. "Mosaicism in Health and Disease — Clones Picking up Speed." *Nature Reviews Genetics* 18 (2). Nature Publishing Group: 128–42. doi:10.1038/nrg.2016.145.

Forsberg, Lars A, Chiara Rasi, Niklas Malmqvist, Hanna Davies, Saichand Pasupulati, Geeta Pakalapati, Johanna Sandgren, et al. 2014. "Mosaic Loss of Chromosome Y in Peripheral Blood Is Associated with Shorter Survival and Higher Risk of Cancer." *Nature Genetics* 46 (6): 624–28. doi:10.1038/ng.2966.

Gaugler, Trent, Lambertus Klei, Stephan J Sanders, Corneliu A Bodea, Arthur P Goldberg, Ann B Lee, Milind Mahajan, et al. 2014. "Most Genetic Risk for Autism Resides with Common Variation." *Nature Genetics* 46 (8). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 881–85. doi:10.1038/ng.3039.

Giraldo, Gustavo, Ana M Gómez, Lina Mora, Fernando Suarez-Obando, and Olga Moreno. 2016. "Mosaic Trisomy 8 Detected by Fibroblasts Cultured of Skin." *Colombia Medica (Cali, Colombia)* 47 (2). Universidad del Valle: 100–104. http://www.ncbi.nlm.nih.gov/pubmed/27546932.

González, Juan R., Alejandro Cáceres, Tonu Esko, Ivon Cuscó, Marta Puig, Mikel Esnaola, Judith Reina, et al. 2014. "A Common 16p11.2 Inversion Underlies the Joint Susceptibility to Asthma and Obesity." *American Journal of Human Genetics* 94 (3): 361–72. doi:10.1016/j.ajhg.2014.01.015.

González, Juan R, Benjamín Rodríguez-Santiago, Alejandro Cáceres, Roger Pique-Regi, Nathaniel Rothman, Stephen J Chanock, Lluís Armengol, and Luis A Pérez-Jurado. 2011. "A Fast and Accurate Method to Detect Allelic Genomic Imbalances Underlying Mosaic Rearrangements Using SNP Array Data." *BMC Bioinformatics* 12 (1): 166. doi:10.1186/1471-2105-12-166.

Gottlieb, Bruce, Lenore K. Beitel, and Mark A. Trifiro. 2001. "Somatic Mosaicism and Variable Expressivity." *Trends in Genetics*. doi:10.1016/S0168-9525(00)02178-8.

Grati, Francesca. 2014. "Chromosomal Mosaicism in Human Feto-Placental Development: Implications for Prenatal Diagnosis."

Journal of Clinical Medicine 3 (3): 809–37. doi:10.3390/jcm3030809.

Gutiérrez-Arumí, Armand. 2016. "Ancestral Genomic Submicroscopic Inversions of Human Genome and Their Relation with Multifactorial Human Diseases." Universitat Pompeu Fabra.

Haitjema, Saskia, Daniel Kofink, Jessica van Setten, Sander W. van der Laan, Arjan H. Schoneveld, James Eales, Maciej Tomaszewski, et al. 2017. "Loss of Y Chromosome in Blood Is Associated With Major Cardiovascular Events During Follow-Up in Men After Carotid EndarterectomyCLINICAL PERSPECTIVE." *Circulation: Cardiovascular Genetics* 10 (4): e001544. doi:10.1161/CIRCGENETICS.116.001544.

Hassold, T, and P Hunt. 2001. "To Err (Meiotically) Is Human: The Genesis of Human Aneuploidy." *Nat Rev Genet* 2 (4): 280–91. doi:10.1038/35066065.

Hastings, P J, James R Lupski, Susan M Rosenberg, and Grzegorz Ira. 2009. "Mechanisms of Change in Gene Copy Number." *Nature Reviews. Genetics* 10 (8): 551–64. doi:10.1038/nrg2593.

Hennes, Jason L., and Meghan Rodes. 2011. *Diagnostic and Statistical Manual of Mental Disorders and Pain Management*. *Essentials of Pain Medicine*. doi:10.1016/B978-1-4377-2242-0.00016-X.

Hu, W F, M H Chahrour, and C A Walsh. 2014. "The Diverse Genetic Landscape of Neurodevelopmental Disorders." *Annu Rev Genomics Hum Genet* 15 (July): 195–213. doi:10.1146/annurev-genom-090413-025600.

Huguet, Guillaume, Elodie Ey, and Thomas Bourgeron. 2013. "The Genetic Landscapes of Autism Spectrum Disorders." *Annual Review of Genomics and Human Genetics* 14 (1). Annual Reviews: 191–213. doi:10.1146/annurev-genom-091212-153431.

Illumina. 2010. "Interpreting Infinium Assay Data for Whole-Genome Structural Variation." *Analysis*, 0–9.

Itsara, Andy, Hao Wu, Joshua D. Smith, Deborah A. Nickerson, Isabelle Romieu, Stephanie J. London, and Evan E. Eichler.

2010. “De Novo Rates and Selection of Large Copy Number Variation.” *Genome Research* 20 (11): 1469–81. doi:10.1101/gr.107680.110.

Jacobs, Kevin B, Meredith Yeager, Weiyin Zhou, Sholom Wacholder, Zhaoming Wang, Benjamin Rodriguez-Santiago, Amy Hutchinson, et al. 2012. “Detectable Clonal Mosaicism and Its Relationship to Aging and Cancer.” *Nature Genetics* 44 (6). Nature Publishing Group: 651–58. doi:10.1038/ng.2270.

Jamuar, Saumya S., Anh-Thu N. Lam, Martin Kircher, Alissa M. D’Gama, Jian Wang, Brenda J. Barry, Xiaochang Zhang, et al. 2014. “Somatic Mutations in Cerebral Cortical Malformations.” *New England Journal of Medicine* 371 (8): 733–43. doi:10.1056/NEJMoa1314432.

Ju, Young Seok, Inigo Martincorena, Moritz Gerstung, Mia Petljak, Ludmil B Alexandrov, Raheleh Rahbari, David C Wedge, et al. 2017. “Somatic Mutations Reveal Asymmetric Cellular Dynamics in the Early Human Embryo.” *Nature* 543 (7647): 714–18. doi:10.1038/nature21703.

Kidd, Jeffrey M., Gregory M. Cooper, William F. Donahue, Hillary S. Hayden, Nick Sampas, Tina Graves, Nancy Hansen, et al. 2008. “Mapping and Sequencing of Structural Variation from Eight Human Genomes.” *Nature* 453 (7191): 56–64. doi:10.1038/nature06862.

King, Bryan H. 2016. “Psychiatric Comorbidities in Neurodevelopmental Disorders.” *Current Opinion in Neurology* 29 (2): 113–17. doi:10.1097/WCO.0000000000000299.

King, Daniel A., Wendy D. Jones, Yanick J. Crow, Anna F. Dominiczak, Nicola A. Foster, Tom R. Gaunt, Jade Harris, et al. 2015. “Mosaic Structural Variation in Children with Developmental Disorders.” *Human Molecular Genetics* 24 (10): 2733–45. doi:10.1093/hmg/ddv033.

Kirkpatrick, Mark. 2010. “How and Why Chromosome Inversions Evolve.” *PLoS Biology* 8 (9). doi:10.1371/journal.pbio.1000501.

Kirkpatrick, Mark, and Nick Barton. 2006. “Chromosome Inversions, Local Adaptation and Speciation.” *Genetics* 173

(1): 419–34. doi:10.1534/genetics.105.047985.

Lapunzina, Pablo, and David Monk. 2011. “The Consequences of Uniparental Disomy and Copy Number Neutral Loss-of-Heterozygosity during Human Development and Cancer.” *Biology of the Cell / under the Auspices of the European Cell Biology Organization* 103: 303–17. doi:10.1042/BC20110013.

Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. 2014. “Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts.” *Genome Biology* 15 (2): R29. doi:10.1186/gb-2014-15-2-r29.

Lim, Elaine T, Mohammed Uddin, Silvia De Rubeis, Yingleong Chan, Anne S Kamumbu, Xiaochang Zhang, Alissa M D’Gama, et al. 2017. “Rates, Distribution and Implications of Postzygotic Mosaic Mutations in Autism Spectrum Disorder.” *Nature Neuroscience*, July. doi:10.1038/nn.4598.

Lo, Min-Tzu, David A Hinds, Joyce Y Tung, Carol Franz, Chun-Chieh Fan, Yunpeng Wang, Olav B Smeland, et al. 2017. “Genome-Wide Analyses for Personality Traits Identify Six Genomic Loci and Show Correlations with Psychiatric Disorders.” *Nature Genetics* 49 (1). NIH Public Access: 152–56. doi:10.1038/ng.3736.

Loomes, Rachel, Laura Hull, and William Polmear Locke Mandy. 2017. “What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis.” *Journal of the American Academy of Child & Adolescent Psychiatry* 56 (6): 466–74. doi:10.1016/j.jaac.2017.03.013.

Lynch, Michael. 2010. “Rate, Molecular Spectrum, and Consequences of Human Mutation.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (3). National Academy of Sciences: 961–68. doi:10.1073/pnas.0912629107.

Machiela, Mitchell J., Weiyin Zhou, Joshua N. Sampson, Michael C. Dean, Kevin B. Jacobs, Amanda Black, Louise A. Brinton, et al. 2015. “Characterization of Large Structural Genetic Mosaicism in Human Autosomes.” *American Journal of Human Genetics* 96 (3): 487–97. doi:10.1016/j.ajhg.2015.01.011.

Marshall, Christian R., Abdul Noor, John B. Vincent, Anath C. Lionel, Lars Feuk, Jennifer Skaug, Mary Shago, et al. 2008. "Structural Variation of Chromosomes in Autism Spectrum Disorder." *American Journal of Human Genetics* 82 (2): 477–88. doi:10.1016/j.ajhg.2007.12.009.

Martínez-Fundichely, Alexander, Sònia Casillas, Raquel Egea, Miquel Ràmia, Antonio Barbadilla, Lorena Pantano, Marta Puig, and Mario Cáceres. 2014. "InvFEST, a Database Integrating Information of Polymorphic Inversions in the Human Genome." *Nucleic Acids Research* 42 (D1). doi:10.1093/nar/gkt1122.

Miles, Judith H. 2011. "Autism Spectrum disorders—A Genetics Review." *Genetics in Medicine* 13 (4). Nature Publishing Group: 278–94. doi:10.1097/GIM.0b013e3181ff67ba.

Milholland, Brandon, Xiao Dong, Lei Zhang, Xiaoxiao Hao, Yousin Suh, and Jan Vijg. 2017. "Differences between Germline and Somatic Mutation Rates in Humans and Mice." *Nature Communications* 8 (May): 15183. doi:10.1038/ncomms15183.

Mitchell, Kevin J. 2011. "The Genetics of Neurodevelopmental Disease." *Current Opinion in Neurobiology* 21 (1): 197–203. doi:10.1016/j.conb.2010.08.009.

Moreno, Eduardo, and Christa Rhiner. 2014. "Darwin's Multicellularity: From Neurotrophic Theories and Cell Competition to Fitness Fingerprints." *Current Opinion in Cell Biology*. doi:10.1016/j.ceb.2014.06.011.

Nagaoka, So I., Terry J. Hassold, and Patricia A. Hunt. 2012. "Human Aneuploidy: Mechanisms and New Insights into an Age-Old Problem." *Nature Reviews Genetics* 13 (7): 493–504. doi:10.1038/nrg3245.

Nybo Andersen, A.-M., and S K Urhoj. 2017. "Is Advanced Paternal Age a Health Risk for the Offspring?" *Fertility and Sterility* 107 (2): 312–18. doi:10.1016/j.fertnstert.2016.12.019.

O'Roak, Brian J., Laura Vives, Santhosh Girirajan, Emre Karakoc, Niklas Krumm, Bradley P. Coe, Roie Levy, et al. 2012. "Sporadic Autism Exomes Reveal a Highly Interconnected Protein Network of de Novo Mutations." *Nature* 485 (7397).

Nature Publishing Group: 246–50. doi:10.1038/nature10989.

Oerlemans, Anoek M., Catharina A. Hartman, Barbara Franke, Jan K. Buitelaar, and Nanda N.J. Rommelse. 2016. "Does the Cognitive Architecture of Simplex and Multiplex ASD Families Differ?" *Journal of Autism and Developmental Disorders* 46 (2): 489–501. doi:10.1007/s10803-015-2572-9.

Osborne, L R, M Li, B Pober, D Chitayat, J Bodurtha, A Mandel, T Costa, et al. 2001. "A 1.5 Million-Base Pair Inversion Polymorphism in Families with Williams-Beuren Syndrome." *Nature Genetics* 29 (3): 321–25. doi:10.1038/ng753.

Ostroverkhova, N V, S A Nazarenko, I N Lebedev, A D Cheremnykh, T V Nikitina, and N N Sukhanova. 2002. "[Detection of Aneuploidy in Spontaneous Abortions Using the Comparative Hybridization Method]." *Genetika* 38 (12): 1690–98. http://www.ncbi.nlm.nih.gov/pubmed/12575456.

Pack, Svetlana D., Robert J. Weil, Alexander O. Vortmeyer, Weifen Zeng, Jie Li, Hiroaki Okamoto, Makoto Furuta, et al. 2005. "Individual Adult Human Neurons Display Aneuploidy: Detection by Fluorescence in Situ Hybridization and Single Neuron PCR." *Cell Cycle* 4 (12): 1758–60. doi:10.4161/cc.4.12.2153.

Pang, Andy W, Jeffrey R MacDonald, Dalila Pinto, John Wei, Muhammad A Rafiq, Donald F Conrad, Hansoo Park, et al. 2010. "Towards a Comprehensive Structural Variation Map of an Individual Human Genome." *Genome Biology* 11 (5): R52. doi:10.1186/gb-2010-11-5-r52.

Persani, Luca, Marco Bonomi, Ana Lleo, Simone Pasini, Fabiola Civardi, Ilaria Bianchi, Irene Campi, et al. 2012. "Increased Loss of the Y Chromosome in Peripheral Blood Cells in Male Patients with Autoimmune Thyroiditis." *Journal of Autoimmunity* 38 (2–3). doi:10.1016/j.jaut.2011.11.011.

Pinto, Dalila, Alistair T. Pagnamenta, Lambertus Klei, Richard Anney, Daniele Merico, Regina Regan, Judith Conroy, et al. 2010. "Functional Impact of Global Rare Copy Number Variation in Autism Spectrum Disorders." *Nature* 466 (7304). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 368–72. doi:10.1038/nature09146.

Pique-Regi, Roger, Jordi Monso-Varona, Antonio Ortega, Robert C. Seeger, Timothy J. Triche, and Shahab Asgharzadeh. 2008. "Sparse Representation and Bayesian Detection of Genome Copy Number Alterations from Microarray Data." *Bioinformatics* 24 (3): 309–18. doi:10.1093/bioinformatics/btm601.

Poduri, Annapurna, Gilad D Evrony, Xuyu Cai, and Christopher A Walsh. 2013. "Somatic Mutation, Genomic Variation, and Neurological Disease." *Science* 341 (6141): 1237758. doi:10.1126/science.1237758.

Puig, Marta, Sònia Casillas, Sergi Villatoro, and Mario Cáceres. 2015. "Human Inversions and Their Functional Consequences." *Briefings in Functional Genomics* 14 (5): 369–79. doi:10.1093/bfgp/elv020.

Puig, Marta, David Castellano, Lorena Pantano, Carla Giner-Delgado, David Izquierdo, Magdalena Gayà-Vidal, José Ignacio Lucas-Lledó, et al. 2015. "Functional Impact and Evolution of a Novel Human Polymorphic Inversion That Disrupts a Gene and Creates a Fusion Transcript." *PLoS Genetics* 11 (10). doi:10.1371/journal.pgen.1005495.

Richardson, Sandra R, Santiago Morell, and Geoffrey J Faulkner. 2014. "L1 Retrotransposons and Somatic Mosaicism in the Brain." *Annual Review of Genetics* 48 (1): 1–27. doi:10.1146/annurev-genet-120213-092412.

Robinson, E. B., P. Lichtenstein, H. Anckarsater, F. Happe, and A. Ronald. 2013. "Examining and Interpreting the Female Protective Effect against Autistic Behavior." *Proceedings of the National Academy of Sciences* 110 (13): 5258–62. doi:10.1073/pnas.1211070110.

Robinson, Wendy P. 2000. "Mechanisms Leading to Uniparental Disomy and Their Clinical Consequences." *BioEssays* 22 (5): 452–59. doi:10.1002/(SICI)1521-1878(200005)22:5<452::AID-BIES7>3.0.CO;2-K.

Rodríguez-Santiago, Benjamín, Núria Malats, Nathaniel Rothman, Lluís Armengol, Montse Garcia-Closas, Manolis Kogevinas, Olaya Villa, et al. 2010. "Mosaic Uniparental Disomies and Aneuploidies as Large Structural Variants of the Human Genome." *American Journal of Human Genetics* 87 (1).

Elsevier: 129–38. doi:10.1016/j.ajhg.2010.06.002.

Rosti, Rasim O., Abdelrahim A. Sadek, Keith K. Vaux, and Joseph G. Gleeson. 2014. “The Genetic Landscape of Autism Spectrum Disorders.” *Developmental Medicine and Child Neurology* 56 (1): 12–18. doi:10.1111/dmcn.12278.

Sachdev, Nidhee M., Susan M. Maxwell, Andria G. Besser, and James A. Grifo. 2017. “Diagnosis and Clinical Management of Embryonic Mosaicism.” *Fertility and Sterility* 107 (1): 6–11. doi:10.1016/j.fertnstert.2016.10.006.

Sakofsky, Cynthia J., and Anna Malkova. 2017. “Break Induced Replication in Eukaryotes: Mechanisms, Functions, and Consequences.” *Critical Reviews in Biochemistry and Molecular Biology*, 1–19. doi:10.1080/10409238.2017.1314444.

Salm, Maximilian P A, Stuart D Horswell, Claire E Hutchison, Helen E Speedy, Xia Yang, Liming Liang, Eric E Schadt, et al. 2012. “The Origin, Global Distribution, and Functional Impact of the Human 8p23 Inversion Polymorphism.” *Genome Research* 22 (6). Cold Spring Harbor Laboratory Press: 1144–53. doi:10.1101/gr.126037.111.

Sanders, Ashley D., Mark Hills, David Porubský, Victor Guryev, Ester Falconer, and Peter M. Lansdorp. 2016. “Characterizing Polymorphic Inversions in Human Genomes by Single-Cell Sequencing.” *Genome Research* 26 (11): 1575–87. doi:10.1101/gr.201160.115.

Sanders, Stephan J., A. Gulhan Ercan-Sencicek, Vanessa Hus, Rui Luo, Michael T. Murtha, Daniel Moreno-De-Luca, Su H. Chu, et al. 2011. “Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism.” *Neuron* 70 (5): 863–85. doi:10.1016/j.neuron.2011.05.002.

Sanders, Stephan J., Xin He, A. Jeremy Willsey, A. Gulhan Ercan-Sencicek, Kaitlin E. Samocha, A. Ercument Cicek, Michael T. Murtha, et al. 2015. “Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci.” *Neuron* 87 (6). Elsevier Inc.: 1215–33. doi:10.1016/j.neuron.2015.09.016.

Sandin, Sven, Paul Lichtenstein, Ralf Kuja-Halkola, Henrik Larsson, Christina M Hultman, and Abraham Reichenberg. 2014. “The Familial Risk of Autism.” *JAMA* 311 (17): 1770–77. doi:10.1001/jama.2014.4144.

Scott, Stuart A, Ninette Cohen, Tracy Brandt, Gokce Toruner, Robert J Desnick, and Lisa Edelmann. 2010. “Detection of Low-Level Mosaicism and Placental Mosaicism by Oligonucleotide Array Comparative Genomic Hybridization.” *Genetics in Medicine* 12 (2). Nature Publishing Group: 85–92. doi:10.1097/GIM.0b013e3181cc75d0.

Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, et al. 2007. “Strong Association of De Novo Copy Number Mutations with Autism.” *Science* 316 (5823): 445–49. doi:10.1126/science.1138659.

Skaletsky, Helen, Tomoko Kuroda-Kawaguchi, Patrick J. Minx, Holland S. Cordum, LaDeana Hillier, Laura G. Brown, Sjoerd Repping, et al. 2003. “The Male-Specific Region of the Human Y Chromosome Is a Mosaic of Discrete Sequence Classes.” *Nature* 423 (6942): 825–37. doi:10.1038/nature01722.

Thompson, Sarah L., and Duane A. Compton. 2008. “Examining the Link between Chromosomal Instability and Aneuploidy in Human Cells.” *Journal of Cell Biology* 180 (4): 665–72. doi:10.1083/jcb.200712029.

U.S. Department of Health and Human Services. 2014. “Prevalence of Autism Spectrum Disorder among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2010.” *Morbidity and Mortality Weekly Report. Surveillance Summaries (Washington, D.C. : 2002)* 63 (2): 1–21. doi:24670961.

Valiente, Manuel, and Oscar Marín. 2010. “Neuronal Migration Mechanisms in Development and Disease.” *Current Opinion in Neurobiology*. doi:10.1016/j.conb.2009.12.003.

van Loo, K M J, and G J M Martens. 2007. “Genetic and Environmental Factors in Complex Neurodevelopmental Disorders.” *Current Genomics* 8 (7). Bentham Science Publishers: 429–44. doi:10.2174/138920207783591717.

Vanneste, Evelyne, Thierry Voet, Cédric Le Caignec, Michèle

Ampe, Peter Konings, Cindy Melotte, Sophie Debrock, et al. 2009. “Chromosome Instability Is Common in Human Cleavage-Stage Embryos.” *Nature Medicine* 15 (5): 577–83. doi:10.1038/nm.1924.

Webster, Alexandre, and Melina Schuh. 2017. “Mechanisms of Aneuploidy in Human Eggs.” *Trends in Cell Biology*. doi:10.1016/j.tcb.2016.09.002.

Winham, Stacey J., Mariza de Andrade, and Virginia M. Miller. 2014. “Genetics of Cardiovascular Disease: Importance of Sex and Ethnicity.” *Atherosclerosis* 241 (1): 219–28. doi:10.1016/j.atherosclerosis.2015.03.021.

Yang, Jian, Naomi R. Wray, and Peter M. Visscher. 2010. “Comparing Apples and Oranges: Equating the Power of Case-Control and Quantitative Trait Association Studies.” *Genetic Epidemiology* 34 (3): 254–57. doi:10.1002/gepi.20456.

Youssoufian, Hagop, and Reed E. Pyeritz. 2002. “Mechanisms and Consequences of Somatic Mosaicism in Humans.” *Nature Reviews Genetics* 3 (10): 748–58. doi:10.1038/nrg906.

Yuen, Ryan K C, Bhooma Thiruvahindrapuram, Daniele Merico, Susan Walker, Kristiina Tammimies, Ny Hoang, Christina Chrysler, et al. 2015. “Whole-Genome Sequencing of Quartet Families with Autism Spectrum Disorder.” *Nature Medicine* 21 (2): 185–91. doi:10.1038/nm.3792.

Yurov, Yuri B., Ivan Y. Iourov, Svetlana G. Vorsanova, Thomas Liehr, Alexei D. Kolotii, Sergei I. Kutsev, Franck Pellestor, et al. 2007. “Aneuploidy and Confined Chromosomal Mosaicism in the Developing Human Brain.” *PLoS ONE* 2 (6): e558. doi:10.1371/journal.pone.0000558.

Zhou, Weiyin, Mitchell J Machiela, Neal D Freedman, Nathaniel Rothman, Nuria Malats, Casey Dagnall, Neil Caporaso, et al. 2016. “Mosaic Loss of Chromosome Y Is Associated with Common Variation near TCL1A.” *Nature Genetics* 48 (5): 563–68. doi:10.1038/ng.3545.

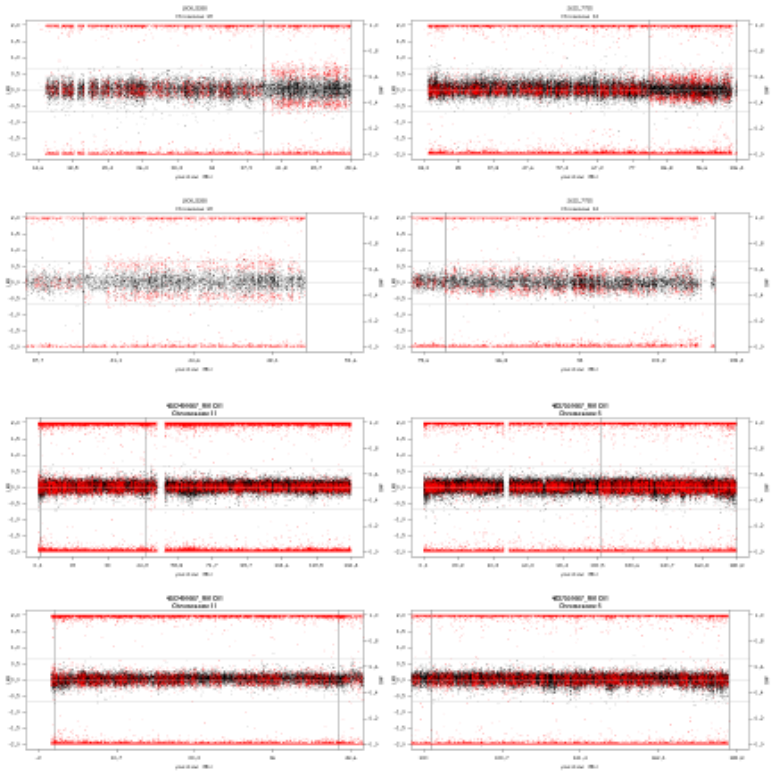
ANNEX

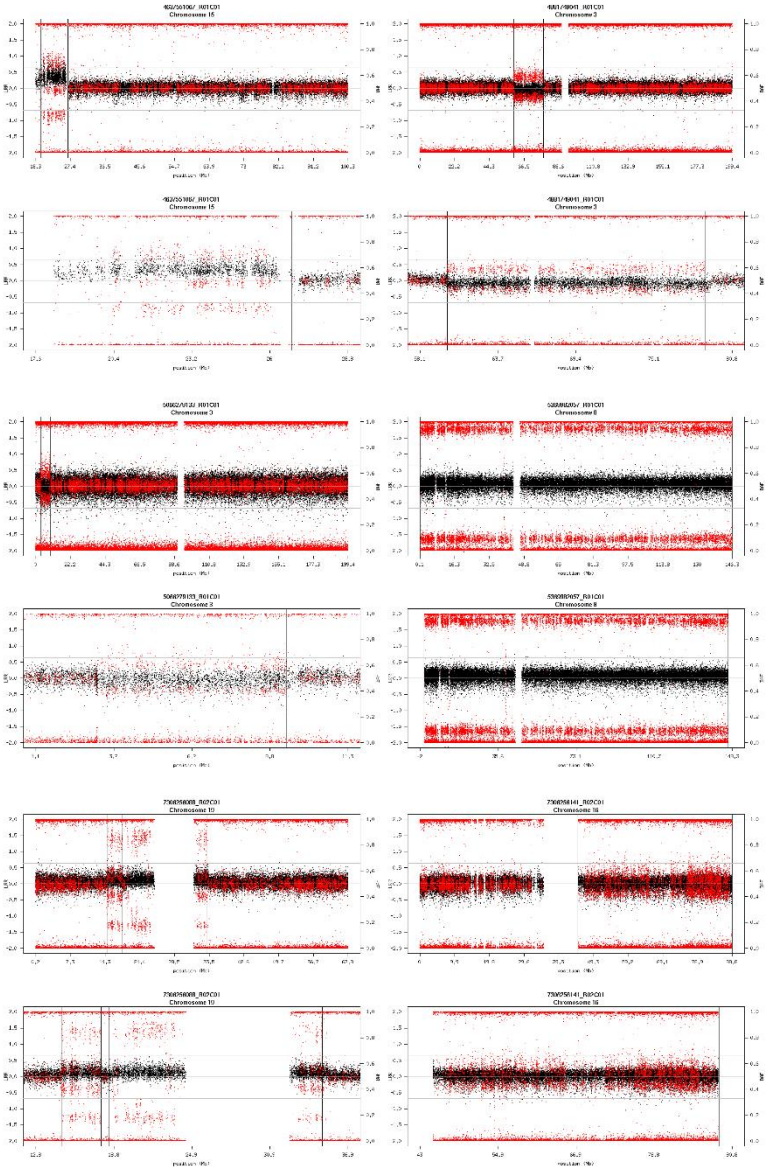
1 Supplementary Material CHAPTER 1

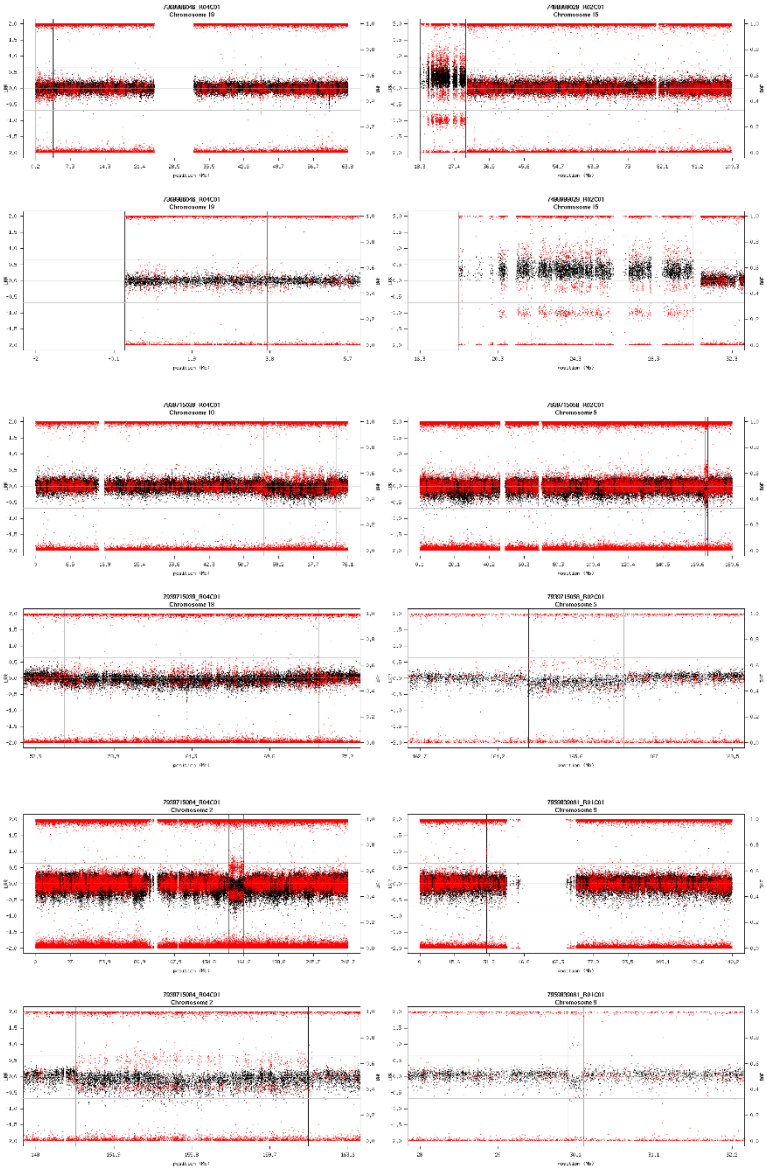
1.1 Supplementary Figures

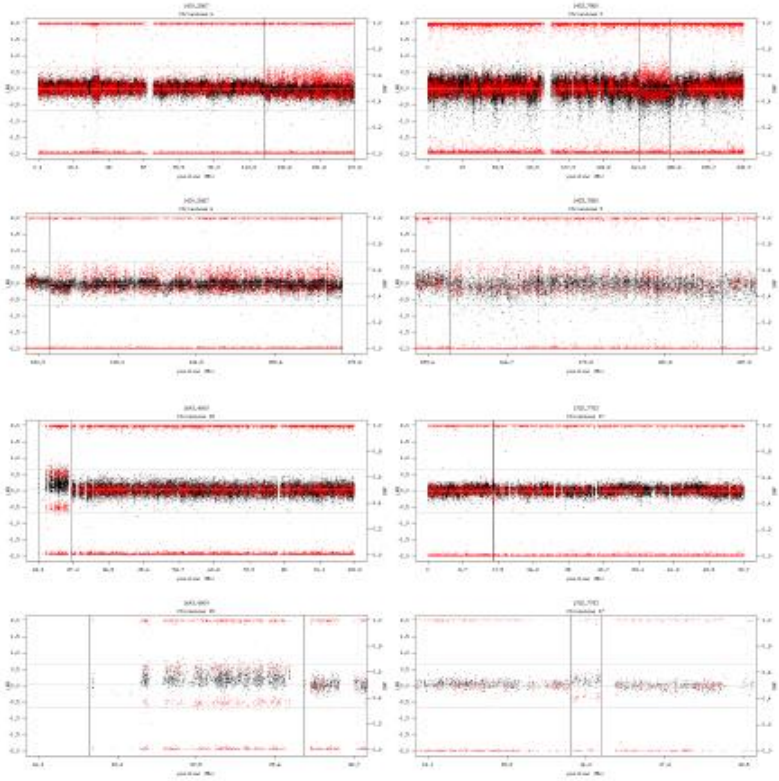
Supplementary Figures 1
Plots of the chromosomal mosaic events detected in blood. For each event there is a view of the whole chromosome, and below there is a zoom to the alteration detected with a 1Mb padding at each breakpoint. The black dots in all plots represent the LRR values for a given probe analysed, while red dots represent the BAF values.

Autosomal Events

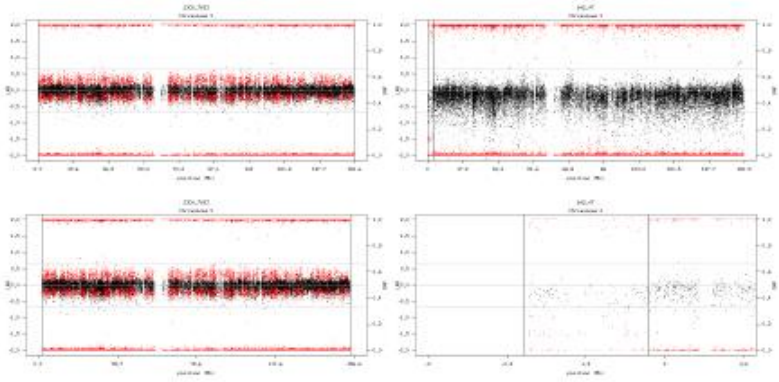




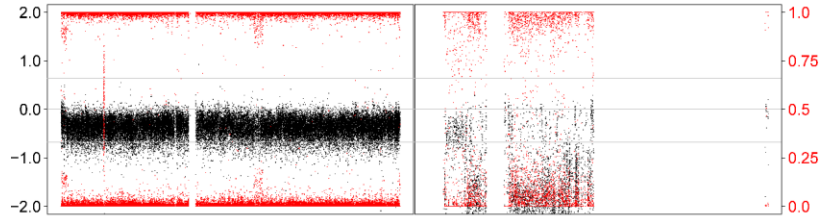




Gonosomal events



Sample 2477_6778 Chromosome X (left) and Chromosome Y (righe)



1.2 Supplementary Tables

Supplementary Table 2
List of events detected in blood samples belonging to parents and siblings.

Sample ID	Role	Study	CMEs (size in Mb)	Start (Mb) (hg18)	End (bp) (hg18)	Cell fraction (%)	Age	Previously reported
1227_4352	Parent	AGP	Loss 11p (3.4)	pter	3,6	39	22 to 50	Yes (1,3)
12514.fa	Parent	SSC	Loss 13q (1.5)	49,7	50,6	41	47	No
1442_3798	Parent	AGP	Loss 20q (12.2)	30,8	43,0	14	22 to 50	Yes (1,3)
12347.mo	Parent	SSC	Loss 20q (1.7)	48,3	50,0	34	41	No
12793.fa	Parent	SSC	Loss 2p (2.1)	24,3	26,3	23	47	No
14166.mo	Parent	SSC	Loss 3p (1.5)	81,9	83,4	59	41	Yes (1,3)
14339.fa	Parent	SSC	Loss 7q (15.5)	111,5	127,0	28	NA	No
11011.fa	Parent	SSC	Loss Yq (3.8)	88,3	92,2	30	60	No
1609_4915	Parent	AGP	Gain 8p (10.6)	36,6	cen	68	22 to 50	No
22_3109	Parent	AGP	Loss Y (57.8)	pter	qter	24	22 to 50	Yes (1,3)
11230.mo	Parent	SSC	Loss X (154.9)	pter	qter	19	40	No
13007.mo	Parent	SSC	Loss X (154.9)	pter	qter	25	56	Yes (1,3)
11116.mo	Parent	SSC	Loss X (154.9)	pter	qter	48	44	No
14556.mo	Parent	SSC	Loss X (154.9)	pter	qter	33	43	No
2132_5873	Parent	AGP	Gain 12 (132.3)	pter	qter	70	22 to 50	No
971_3324	Parent	AGP	Gain 12 (132.3)	pter	qter	35	22 to 50	No
12845.mo	Parent	SSC	Gain 12 (132.3)	pter	qter	27	41	No
14528.fa	Parent	SSC	Gain 21 (37.2)	pter	qter	34	43	No
1339_3174	Parent	AGP	UPD 11p (40.9)	pter	41,0	26	22 to 50	No
12750.fa	Parent	SSC	UPD 15q (58.7)	41,6	qter	20	44	No
1648_6719	Parent	AGP	UPD 17q (56.3)	cen	qter	16	22 to 50	No
11168.mo	Parent	SSC	UPD 17p (21.9)	pter	cen	37	44	No
11804.fa	Parent	SSC	UPD 17q (42.7)	36,0	qter	6	43	No
14649.mo	Parent	SSC	UPD 17q (1.0)	77,6	qter	12	NA	No
11055.fa	Parent	SSC	UPD 1q (101.8)	cen	qter	9	58	No
13183.fa	Parent	SSC	UPD 1q (105.7)	cen	qter	8	38	No
19_1047	Parent	AGP	UPD 1p (108.8)	pter	109,6	10	22 to 50	No
12421.mo	Parent	SSC	UPD 21q (31.1)	15,8	qter	51	45	No
384_3197	Parent	AGP	UPD 22q (32.5)	17,0	qter	22	22 to 50	No
13365.mo	Parent	SSC	UPD 2q (8.5)	234,2	qter	12	32	No
1451_960	Parent	AGP	UPD 3p (17.7)	pter	17,8	10	22 to 50	No
577_4090	Parent	AGP	UPD 5q (8.6)	172,1	qter	13	22 to 50	No
13418.fa	Parent	SSC	UPD 9p (28.1)	pter	28,1	57	54	No
1194_3356	Parent	AGP	UPD 9q (70.5)	69,6	qter	62	22 to 50	No
1942_5813	Parent	AGP	UPD 9q (69.6)	70,5	qter	12	22 to 50	No
14266.s1	Sibling	SSC	Loss 14q (0.5)	63,5	63,9	54	11	No
			Loss Xp (56.4)	pter	cen	54		No
			Gain Xq (98.4)	cen	qter	54		No
13391.s1	Sibling	SSC	Loss 4q (3.2)	141,0	144,1	33	6	No
			Gain 4q (11.5)	126,0	137,5			No
14158.s1	Sibling	SSC	UPD 13 (96.2)	pter	qter	9	14	No

UPD: Uniparental isodisomy
(1) Pinto et al 2010
(2) Sanders et al 2011
(3) Pinto et al 2014
(4) Sanders et al 2015

Supplementary Table 3
List of events detected in blood belonging to ASD patients smaller than 0.4 Mb

Sample ID	Study	CNEs (size in Mb)	Start (Mb)	End (Mb)	Cell fraction (%)	Age at DNA Sampling	Sex	Simplex / Multiplex	Diagnosis/ADI-R/ADOS	Verbal	Intelectual Disability	Genes in region	Genes adjacent to region	Previously reported
11270.p1	SSC	Loss 9p22 (0.12)	12,8	12,9	27	6	Male	Simplex	Autism	Yes	Yes	C9orf79	DLC1	No
13603.p1	SSC	Loss 9p21 (0.21)	30,0	30,2	37	4	Male	Simplex	ASD-Autism	Yes	Yes	-	LINGO2	No

Supplementary Table 4
ASD probands reported Copy-number variants that are located or overlap with the breakpoints of the chromosomal mosaic events detected. Reported by Sanders et al 2011.

		Reported by Sanders et al 2011 or Sanders et al 2015							
Sample ID	Study	Band	Chr	Start (hg18)	Stop (hg18)	Size (Kb)	Del/Dup	Parent of Origin	Mechanism
1483_7880	AGP								
1420_2867	AGP								
1643_4809	AGP								
1782_7783	AGP								
2632_7755	AGP								
1808_5288	AGP								
2300_7693	AGP								
642_47	AGP								
2477_6778	AGP								
11671.p1	SSC								
11270.p1	SSC								
13362.p1	SSC	2q33.1	2	201785316	201839832	0.55	Dup	Father	HR
11238.p1	SSC								
12246.p1	SSC								
14687.p1	SSC	15q11.1-13.3	15	18265295	30407419	121.42	Dup	Mother	NAHR
13006.p1	SSC								
11679.p1	SSC								
14466.p1	SSC								
12245.p1	SSC	22q11.22	22	21100917	21567586	4.67	Del	Unknown	NHR
12007.p1	SSC	15q11.1-13.1	15	18275409	26728046	84.53	Dup	Mother	NAHR
14556.p1	SSC	19p12	19	20867601	20944207	0.77	Dup	Unknown	NHR

HR: Homologous Recombination
NAHR: Non-allelic Homologous Recombination

Supplementary Table 5
ASD probands reported variants that are located inside the breakpoints of the chromosomal mosaic events detected. Reported by Iossifov et al 2014.

		Reported by Iossifov 2014		
Sample ID	Study	Location	effectGene	effectType
1483_7880	AGP			
1420_2867	AGP			
1643_4809	AGP			
1782_7783	AGP			
2632_7755	AGP			
1808_5288	AGP			
2300_7693	AGP			
642_47	AGP			
2477_6778	AGP			
11671.p1	SSC			
11270.p1	SSC			
13362.p1	SSC			
11238.p1	SSC			
12246.p1	SSC			
14687.p1	SSC	13:51948834:G:A	INTS6	nonsense
		16:28900149:C:A	ATP2A1	missense
13006.p1	SSC			
11679.p1	SSC	2:209150500:C:T	PIKFYVE	missense
14466.p1	SSC			
12245.p1	SSC			
12007.p1	SSC			
14556.p1	SSC			

2 Supplementary Material CHAPTER 2

2.1 Online Methods

Online Methods

Robust estimation of mosaic loss of chromosome Y with genotype-array-intensity data

Juan R Gonzalez[1,2,3,4], Marcos Lopez[1,3], Pedro Puig[4], Tonu Esko[5], Luis A Perez-Jurado[3]

Statistical Methods

LRR and mLRR-Y summarization

Raw data must be processed to obtain required data in PennCNV format (http://penncnv.openbioinformatics.org/en/latest/user-guide/input/). Basically, several files of each sample must be created containing information about SNP, chromosome, position, LRR, BAF and genotype (although having only the first 4 columns is enough to call LOY events). Different tools can be used to get the required information. Affymetrix data (CEL files) can be processed by using Birdseed v2 algorithm (http://archive.broadinstitute.org/mpg/birdsuite/birdseed.html). The algorithm provides the copy-number state (CNS) of each probe and the copy-number state of each allele at each probe. The LRR is obtained from the copy-number state for each probe by using the relationship:

$$CNS = Ploidy \cdot \exp(1.5\, LRR)$$

Where *Ploidy* is the copy-number state for the whole individual, usually 2. This formula can be transformed to obtain the expected LRR for a known copy-number state:

$$LRR = \frac{2}{3}\log\left(\frac{CNS}{Ploidy}\right)$$

The BAF were obtained by setting to zero the negative allele copy-number values and then, calculating the relationship between the B allele copy-number estimation versus the sum of both alleles. Affymetrix power tools can also be used to process .CEL files (https://www.affymetrix.com/support/developer/powertools/changelog/index.html) as well as affy2sv R package (https://bitbucket.org/brge/affy2sv/wiki/Home). Illumina data (idat files) can be processed by using Genome Studio software (https://www.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html). crlmm Bioconductor package can also be used to get LRR and BAF (crlmm -https://www.bioconductor.org/packages/release/bioc/html/crlmm.html).

pennCNV files are then processed with MADloy package to summarize the mLRR-Y data (median) and the LLR in a reference region that by default correspond to all autosomes (trimmed-mean).

LRR data normalization

There are several artifacts that have to be taken into account before calling LOY samples using mLRR-Y information. First, there can be a systematic bias in the mLRR-Y due to the fact that the median intensity of LRR distribution shifted slightly away from 0 in the whole array. This

issue is addressed by normalizing the median mLRR-Y data using the a robust median LRR intensity in the autosomes (Cheng et al. 2011). In particular, we propose to compute the 5% trimmed-mean of LRR to avoid regions having copy number alterations. The percentage can be tuned to take into account the different nature of the data we are dealing with. As an example, studies in cancer are expected to have individuals with large number of aneuploidies. Therefore, the trimmed value of the LRR may be increased up to, for instance, 25%. Second, some of the existing algorithms used to get LRR information considers two copies of each chromosome, including chromosome Y, providing LRR values close to -0.46 for chromosome Y, indicating that only 1 copy is present in males. We address this issue by recalculating values of LRR in the msY region, first estimating the real copy-number state of LRR values in this region using the previously introduced relation between *LRR*, *CNS* and *Ploidy*, considering a Ploidy value of 2 for the whole chromosome Y:

$$CNS = 2 \exp(\, 1.5\, LRR\,)$$

After obtaining the copy-number state values, the LRR is recalculated by applying a derivation of the previous formula but considering a *Ploidy* value of 1, that results in the following formula:

$$LRR = \frac{2}{3} \log(\, CNS\,)$$

Quality control

We perform a quality control of the samples involved in the analyses by removing those samples with large LRR variability, just to avoid introducing extra variability due to technical artifacts. We follow the recommendation proposed in this technical report (www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote_cnv_loh.pdf) that proposes to filter those samples having high variable LRR in the reference region. In particular, they consider LRR standard deviation (LRR_SD) larger than 0.28 as bad samples. Our package can consider other filtering options. The user can manually set LRR_SD, or it can be estimated from the data considering as 2 times the mean observed LRR_SD from all samples.

Calling LOY using m-LRR-Y data

Forsberg et al. (2014), Dumanski et al. (2015) and Dumanski et al. (2016) (Forsberg et al. 2014; J. P. Dumanski et al. 2015; Jan P. Dumanski et al. 2016) assume that experimental variation in mLRR-Y is distributed in a non-skewed fashion. Therefore, they propose to generate the expected experimental background noise of mLRR-Y data using a symmetrical distribution. They impose the observed variation in the positive tail of the mLRR-Y distribution into a reflected negative tail by mirroring over median value of mLRR-Y the data observed in the positive tail. They argue that this assumption is reasonable since their validation experiments do not confirm that large values of mLRR-Y are real cases of XYY (e.g. gains of chromosome Y). This assumption may be realistic in some cases; however, it is expected that large values of mLRR-Y indicate the existence of real XYY cases. For instance, by analyzing cases called as

XYY in the TCGA we are able to demonstrate that this assumption does not hold in all situations (see experimental validation section).

In order to overcome these difficulties, we propose to use the trimmed-mean LRR in the autosomes as the data to estimate the real background distribution of the LRR data. **Figure 3 in the Supplementary Material 1** illustrates how the observed empirical LRR distribution is not properly fitted by using a symmetrical distribution (e.g. Gaussian). We observe that the Negative Inverse Gaussian (NIG), a distribution designed to model skewed data, is doing well in all scenarios.

The NIG distribution is a flexible, four parameter distribution that can describe a wide range of shapes. It is just a variance-mean mixture of a Gaussian distribution with an inverse Gaussian. Its density function has the expression,

$$f(x) = \frac{\alpha\delta}{\pi} \frac{e^{\beta(x-\mu)+\delta\sqrt{\alpha^2-\beta^2}}}{\sqrt{(x-\mu)^2+\delta^2}} K_1(\alpha\sqrt{(x-\mu)^2+\delta^2}),$$

where $K_1()$ denotes the modified Bessel function of the third kind with index 1. Each parameter has a different effect on the distribution. α is related to the behavior of the tails, β is the parameter of skewness, μ is the location parameter and δ is the scale parameter measuring the spread of the data.

Therefore, our calling procedure is based on estimating the four parameters of the NIG model based on the LRR data in autosomes, and then assigning a probability of belonging to this null distribution to the mLRR-Y data of each sample. False positive results are controlled by considering that the probability for declaring that an individual is having LOY is 0.05/n where n stands for the number of samples analyzed.

Improving calling including B deviation information

To improve the calling method in those individuals where LRR is not enough accurate to determine whether or not there is a loss, we propose to include the B deviation information in the pseudoautosomal regions PAR1 and PAR2. Usually, the PAR1 and PAR2 regions are not codified as chromosome Y, but as chromosome XY or 25. B deviation is measured in each PAR region by selecting all probes with B Allele Frequency values higher than 0.15 and lower than 0.85 within each region. B deviation is calculated as described in (Rodríguez-Santiago et al. 2010). Those samples with a B deviation > 0.06 and LRR > -0.46 (in case it is normalized with ploidy 2) are reported as possible LOY, and they will have to be verified by visually inspection using functions implemented in MADloy. Alterations of B deviation in PAR regions can also be due to alterations in chromosome X and should be taken into consideration.

Experimental validation

Ten random samples belonging to EGCUT cohort were validated using two Multiplex Ligation Probe-dependent Amplification (MLPA) panels, P070 (MRC-Holland Amsterdam, The Netherlands) covering all subtelomeric regions including the two pseudoautosomal regions (PAR1 and PAR2). Probes targeting Y chromosome, and a custom-made panel with probes for SRY and several autosomal loci, were used to assess the copy number status of chromosome Y with respect to the control loci (autosomal and X chromosome). The MLPA reactions were

carried out essentially as described previously (Schouten et al 2002) with slight modifications when custom probes were used (Rodriguez-Santiago et al 2010b). We used the relative peak height (RPH) method recommended by MRC-Holland. Theoretically, non-mosaic losses and gains show a relative peak height of approximately 0.5 and 1.5, respectively, for the pseudoautosomal regions (normally disomic), and 0 and 2 for the Y-unique regions (normally monosomic).

25 samples for each calling classification belonging to TCGA dataset were randomly selected to be validated using exome data, as proposed in (Forsberg et al. 2014) by computing the median copy-number estimation in the mLRR-y region. Only 58 of the 75 selected samples were available (19 LOY, 18 normal, 21 XYY) and classification of Control-FREEC was compared to the results obtained by our proposed method.

Software

We have created MADloy, a Bioconductor package that automates LOY detection and helps in performing association studies. The core functions include a pipeline to normalized, perform quality control and summarize the processed SNP array data of separate sample files containing information about LRR and BAF in PennCNV format. MADloy contains a function to call LOY and gains of chromosome Y (XYY) using NIG distribution. There are several functions to visualize the calling and the LRR or LRR and BAF in the chromosome Y where mLRR-Y region is highlighted. These plots may help to visually inspect those samples called as LOY. A vignette illustrating how to analyze a subset of samples can be found in https://github.com/isglobal-brge/MADloy/tree/master/vignettes.The development version of this package is available in the GitHub page of our group (https://github.com/isglobal-brge/MADloy).

Data analyses

Separate files for each sample having data in PennCNV format (e.g., SNP name, chromosome, position, LRR, BAF and genotype) are required. We have generated the data of the three different datasets that will be used through this manuscript to illustrate different issues when detecting and analyzing LOY in association studies.

EGCUT general population cohort

First, a total of 126 randomly selected samples (4 females, 122 males) older than 18 years (mean age 51.6 +/- 18.3) from Estonian Gene Expression Cohort (EGCUT, www.biobank.ee). We included 4 males in order to illustrate how our tool is able to filter female samples based on genomic data. EGCUT comprises a large cohort of 53,000 samples of the Estonian Genome Center Biobank, University of Tartu (Mestpalu, 2004). Data were genotyped using HumanCoreExome array and all the individuals included in the analysis had a genotyping success rate above 95%. Cryptic relatedness was tested with the PLINK v1.07 software. Only one of each detected relative pairs (up to second cousins) was randomly chosen for the detection of genetic mosaicim. Sample mix-ups were corrected using MixupMapper (Westra et al., 2011). All studies were performed in accordance with the ethical standards of the responsible committee on human experimentation, and with proper informed consent from all individuals tested. LRR and BAF were generated using GenomeStudio software.

TCGA cancer datasets

688 paired tumor and normal male samples of Kidney Renal Clear Cell Carcinoma (KIRC) and 567 paired tumor and normal male samples diagnosed with Low Grade Glioma (LGG) belonging to TCGA project have been analyzed. Raw data obtained from the Genome-Wide Human SNP Array 6.0 chip were processed with Birdseed v2 algorithm. Clinical data were also downloaded to perform downstream analyses and to select male samples. RNAseq data was obtained from RTCGA.rnaseq Bioconductor package that provides counts of 20532 annotated genes in 36 different tumors.

Alzheimer's disease

Data from National Institute on Aging (NIA) - Late Onset Alzheimer's Disease Family Study available at dbGAP under accession number phs000168.v2.p2 have been obtained from Affymetrix 6.0 array. A total of 644 CEL files were downloaded corresponding to male samples. Raw data were processed with Affymetrix Power Tools that allow applying the same transformation as previously described in the case of TCGA data to get LRR and BAF. Age at diagnosis of dementia was also obtained and was used as our event of interest (n=278).

Statistical analyses

Association with age was performed in EGCUT and TCGA data by using linear regression models. Models for TCGA data were adjusted for cancer status (normal/tumor sample). Cox proportional hazard model was used to assess the association between LOY and the age of diagnosis in the NIA study. Transcriptome RNAseq data analysis was performed using *voom* method from *limma* package (Law et al. 2014)

The file **Supplementary Material 1** contains a reproducible document that can be used to generate all the results that are described in the main manuscript.

Simulation studies

We ran a number of simulations to determine the power of the association analyses considering LOY information as a continuous covariate (e.g. Wright et al. approach), a categorical covariate using threshold method (e.g. Forsberg et al. approach) and a categorical covariate using our calling procedure based on NIG distribution. We consider two main scenarios to assess the performance in the case of having continuous (e.g. age) or categorical (e.g. case/control) outcomes. Data were simulated using functions available in the CNVassoc package. In the continuous case data simulations required information about: mean and standard deviation of mLRR in normal and LOY cases, the proportion of cases having LOY, the mean effect and the standard deviation of the outcome. These parameters were obtained from EGCUT data that mimic real situations when analyzing correlation between age and mLRR (i.e. the relationship between age and mLRR is the same as the one observed in Figure XX from Wright et al.). The effect size varied from 1 up to 8 which can be interpreted as age changes for one-unit change in the LOY variable. Sample size varied as {200, 300, 500, 750, 1000, 1500}. The categorical case data simulation is similar but considering odds ratio (OR) as the effect of LOY. We ran different scenarios varying OR as {1.5, 1.75, 2, 2.5, 3, 3.5, 4}. Data simulation and model parameters are further described in **Supplementary File 2** that is a reproducible document that can be used to generate all simulation results.

Bibliography

Cheng, Jiqiu, Evelyne Vanneste, Peter Konings, Thierry Voet, Joris R Vermeesch, and Yves Moreau. 2011. "Single-Cell Copy Number Variation Detection." *Genome Biology* 12 (8). BioMed Central: R80. doi:10.1186/gb-2011-12-8-r80.

Dumanski, J. P., C. Rasi, M. Lonn, H. Davies, M. Ingelsson, V. Giedraitis, L. Lannfelt, et al. 2015. "Smoking Is Associated with Mosaic Loss of Chromosome Y." *Science* 347 (6217): 81–83. doi:10.1126/science.1262092.

Dumanski, Jan P., Jean Charles Lambert, Chiara Rasi, Vilmantas Giedraitis, Hanna Davies, Benjamin Grenier-Boley, Cecilia M. Lindgren, et al. 2016. "Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease." *American Journal of Human Genetics* 98 (6): 1208–19. doi:10.1016/j.ajhg.2016.05.014.

Forsberg, Lars A, Chiara Rasi, Niklas Malmqvist, Hanna Davies, Saichand Pasupulati, Geeta Pakalapati, Johanna Sandgren, et al. 2014. "Mosaic Loss of Chromosome Y in Peripheral Blood Is Associated with Shorter Survival and Higher Risk of Cancer." *Nature Genetics* 46 (6): 624–28. doi:10.1038/ng.2966.

Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts." *Genome Biology* 15 (2): R29. doi:10.1186/gb-2014-15-2-r29.

Rodríguez-Santiago, Benjamín, Núria Malats, Nathaniel Rothman, Lluís Armengol, Montse Garcia-Closas, Manolis Kogevinas, Olaya Villa, et al. 2010. "Mosaic Uniparental Disomies and Aneuploidies as Large Structural Variants of the Human Genome." *American Journal of Human Genetics* 87 (1). Elsevier: 129–38. doi:10.1016/j.ajhg.2010.06.002.

2.2 Supplementary Material 1

Supplementary Material: Robust estimation of mosaic loss of chromosome Y with genotype-array-intensity data

Juan R Gonzalez, Marcos Lopez, Pere Puig, Tonu Esko, Luis A Perez-Jurado

Contents

1 Getting started
2 EGCUT data
  2.1 Check gender status
  2.2 Age association analysis
3 TCGA data
  3.1 KIRC
  3.2 BLCA
  3.3 Cancer association analysis
  3.4 Age association analysis
4 Alzheimer's disease data
5 Transcriptomic analysis
  5.1 KIRKC data
  5.2 BLCA data
  5.3 Heatmap on chromosome Y genes
6 Session information

1 Getting started

Let us load required packages

```
library(MADloy)
library(lme4)
library(CNVassoc)
library(xtable)
library(limma)
library(sva)
library(biomaRt)
library(pheatmap)
library(RColorBrewer)
```

These are auxiliary functions that are required in data analyses

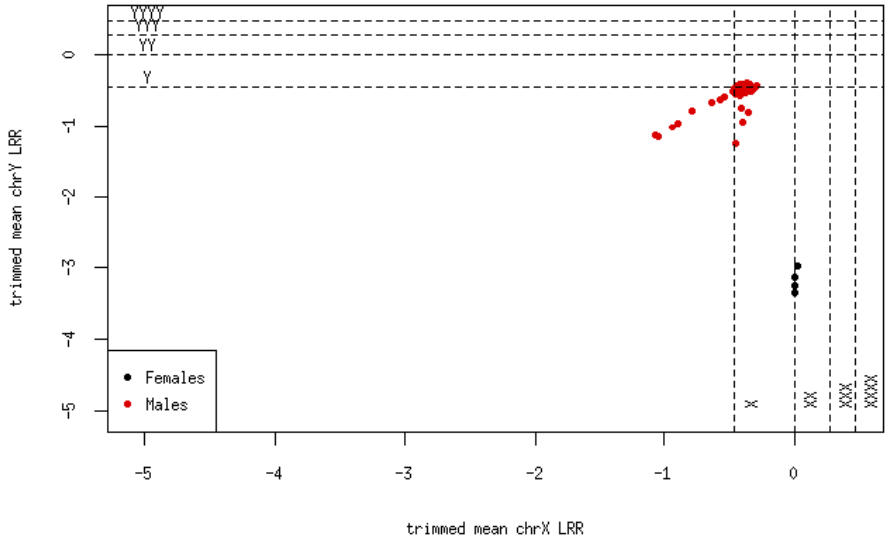
```
de_voom <- function(formula1, formula2, counts, pheno, sva=TRUE) {
 mod <- model.matrix(formula1, data=pheno)
 v <- voom(counts, design=mod)
 if(sva) {
   mod0 <- model.matrix(formula2, data=pheno)
   ns <- num.sv(counts, mod, method="be")
   ss <- svaseq(counts, mod, mod0, n.sv=ns)$sv
   modss <- cbind(mod, ss)
   fit <- lmFit(v, modss)
 }
 else {
   fit <- lmFit(v, mod)
 }
 fit <- eBayes(fit)
 fit
}

getThreshold <- function(x, ...)
{
  den <- density(x[!is.na(x)], bw="SJ", ...)
  m <- den$x[which(den$y==max(den$y))]
  xx <- x[x>=m]
  x.99 <- quantile(xx, 0.99, na.rm=TRUE)
  ans <- m - x.99
  ans
}

GenesRegion <- function(fit, genes, coef=2) {
    tt <- topTable(fit, num = Inf, coef=coef)
    n <- unlist(strsplit(rownames(tt), "\\|.*"))
    qm = which(n == "?" | n == "SLC35E2")
    tt = tt[-qm,]
    n = n[-qm]
    rownames(tt) = n
    i <- intersect(genes, n)
    p = tt[i,]
    return(list("Fit" = fit, "topTable" = tt, "topTableRegion" = p, "i
}
```

2 EGCUT data

2.1 Check gender status

```
path.egcut <- "/SYNCRW10125/DATASETS/STUDY/EGCUT/rawData_anon"
sex.egcut <- checkSex(path.egcut, mc.cores=26)
```

```
plot(sex.egcut)
```




```
egcut.males <- sex.egcut$par$files[sex.egcut$class=="MALE"]
```

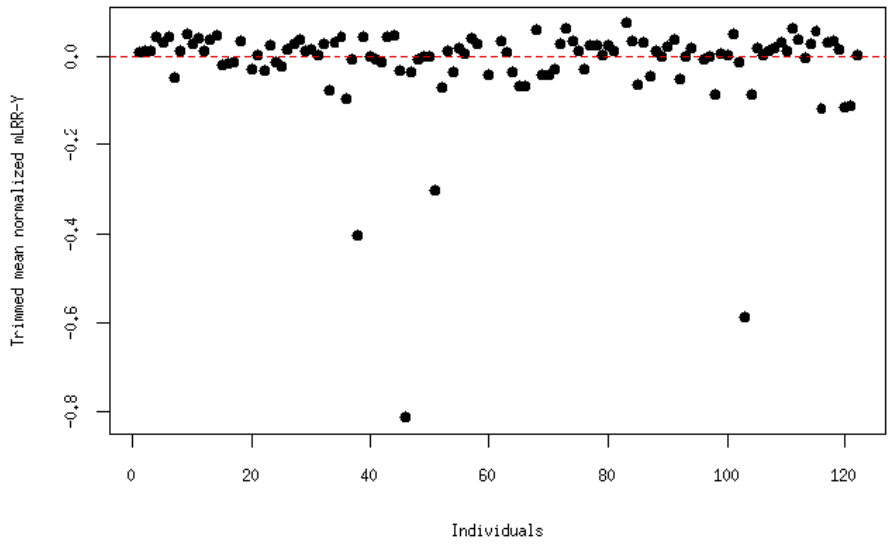
Let us preprocess raw data files to get LRR and BAF information in the mLRR-Y region and the autosomes

```
egcut <- madloy(egcut.males, trim=0.10, mc.cores=26)
egcut
```

```
## Object of class MADloy
## ----------------------------
## Number of processed samples: 122
## Target region: chrY:6671498-22919969
## Reference region(s): Autosomal chromosomes
## Offset (median LRR value in msY): -0.54
```

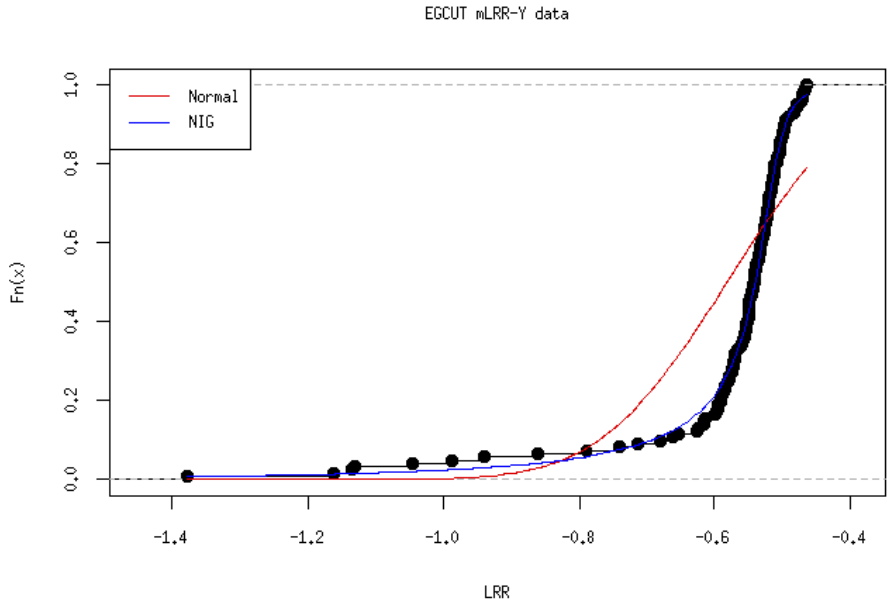
This fiture illustrate the differences with regard to the mLRR-Y among individuals. Those having values around 0 are considered to be normal.

```
plot(egcut, ylim=c(-2, 1))
```



Let us perform calling. It is based on Negative Inverse Gaussian distribution that is properly fitting LRR data as can be seen in this figure

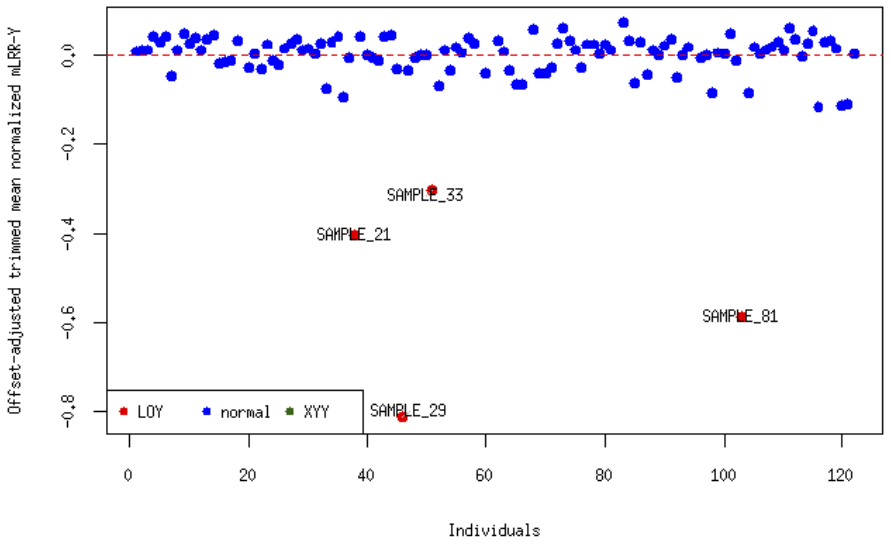
```
plotNIG(egcut, "EGCUT mLRR-Y data")
```



The calling is performed by using `getLOY` function. Notice that by default the `offset` argument of this function corresponds to the median value of mLRR-Y in all individuals. The `pval.sig` argument that is used to control the false discovery rate. By default it is based on Bonferroni correction that is set by default.

```
egcut.call <- getLOY(egcut)
```

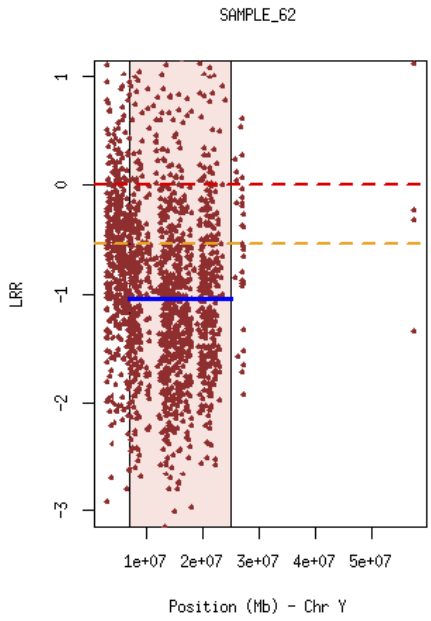
```
plot(egcut.call, ylim=c(-2, 1), print.labels=TRUE)
```

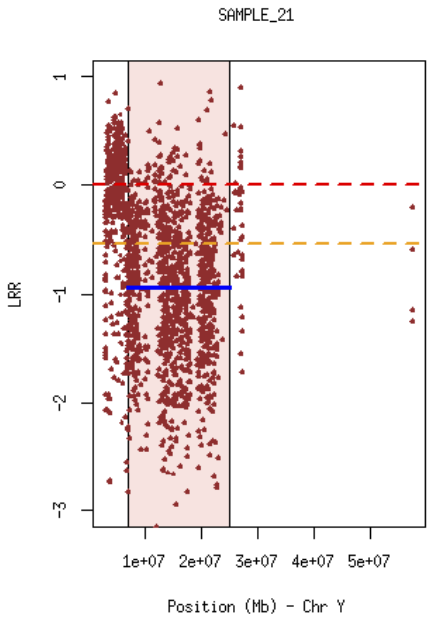


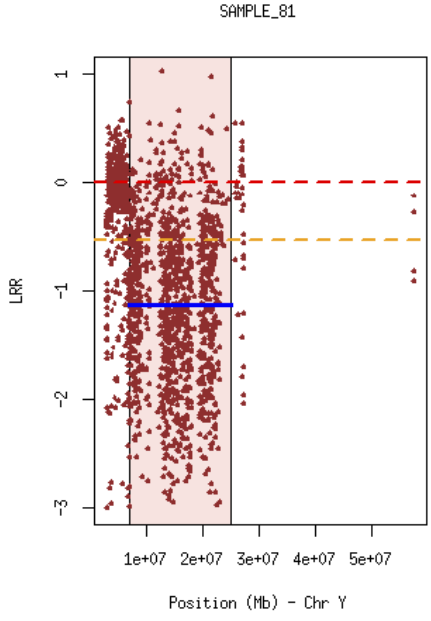

```
par(mfrow=c(2,2))
plotIndLRR(egcut, sample="SAMPLE_62", ylim=c(-3,1))
```

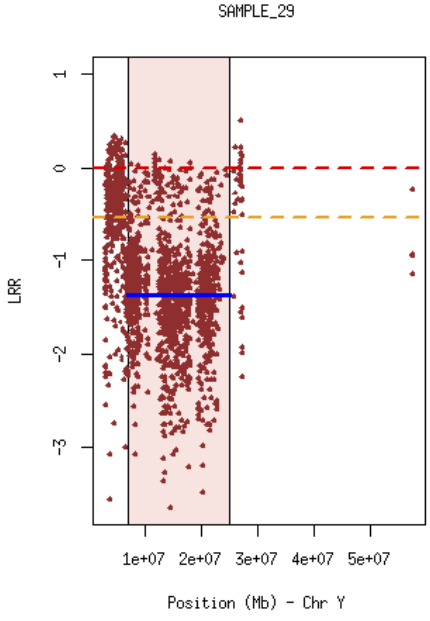
```
##
Read 47.7% of 733202 rows
Read 733202 rows and 5 (of 5) columns from 0.029 GB file in 00:00:03
```

```
plotIndLRR(egcut, sample="SAMPLE_21", ylim=c(-3,1))
plotIndLRR(egcut, sample="SAMPLE_81", ylim=c(-3,1))
plotIndLRR(egcut, sample="SAMPLE_29")
```









2.2 Age association analysis

Let's us perform association analysis between age and the LOY status that has clearly been demonstrated. We are going to use different strategies proposed in the literature.

- Age versus LOY status defined by using our proposed calling method
- Age versus LOY status defined by using the threhold method proposed by Fosbert et al. 2014 and Dumansky et al. 2015
- Age versus LOY by considering mLRR-Y as quantitative trait

```
load("../EGCUT/egcut_age.Rdata")
info <- read.delim("/SYNCRW10125/DATASETS/STUDY/EGCUT/sample_mapping_E
                   as.is=TRUE)
rownames(info) <- paste0(rownames(info), ".txt")
ids <- info[names(egcut.call$class),]
pheno.egcut <- pheno[ids, ]

pheno.egcut$loy <- droplevels(egcut.call$class)
pheno.egcut$loyc <- egcut.call$data
tt <- getThreshold(pheno.egcut$loyc)
pheno.egcut$loyt <- relevel(cut(pheno.egcut$loyc, c(-Inf, tt, Inf),
                                labels=c("LOY", "normal")),2)

mod.egcut <- glm(Age ~ loy, data=pheno.egcut)
mod.egcut.c <- glm(Age ~ loyc, data=pheno.egcut)
mod.egcut.t <- glm(Age ~ loyt, data=pheno.egcut)
```

The results using calling LOY are:

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 49.03571   1.544248 31.753791 1.917199e-58
## loyLOY      25.96429   8.316028  3.122198 2.275312e-03
```

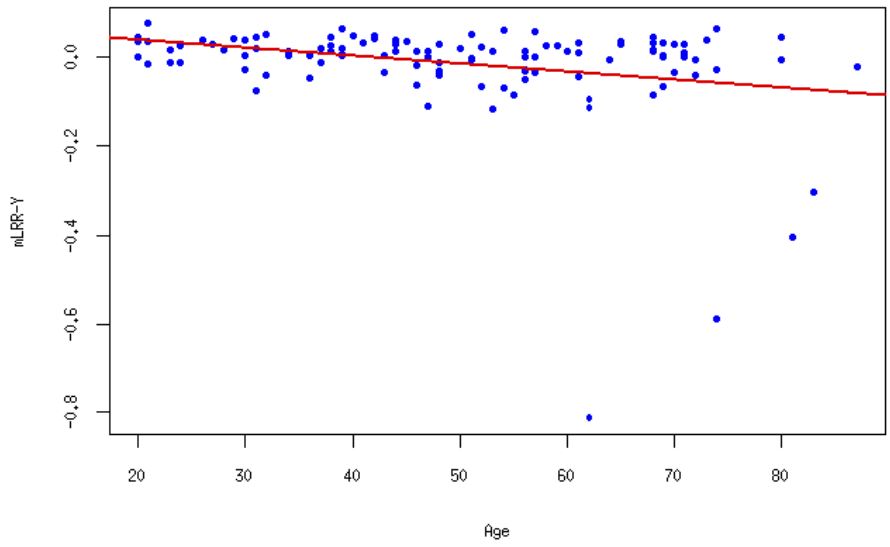
The results using theshold on mLRR-Y are:

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 48.44554   1.649500 29.369829 5.358606e-55
## loytLOY     11.48779   4.587074  2.504382 1.368076e-02
```

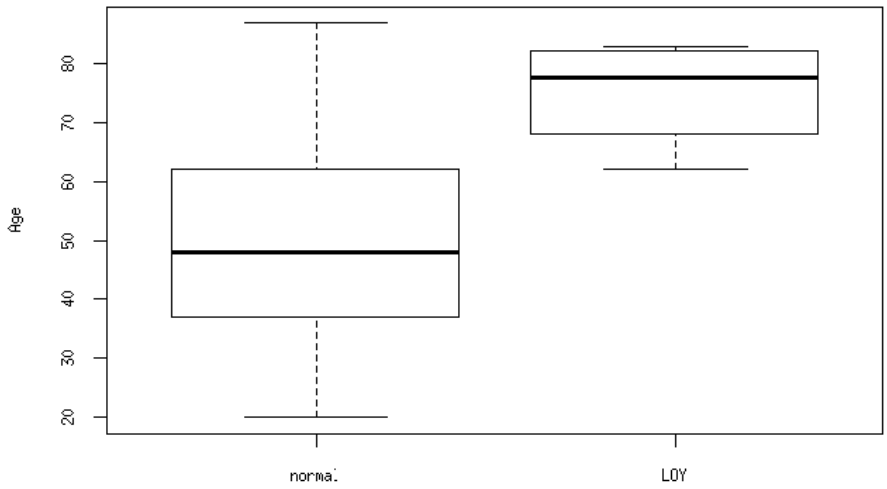
The results considering mLRR-Y as a quantitative trait are:

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  49.31216   1.534215 32.141620 5.514513e-59
## loyc        -41.76791  13.753501 -3.036893 2.963069e-03
```

```
with(pheno.egcut, plot(Age, loyc, xlab="Age", ylab="mLRR-Y", type="n"
                       , cex.axis=1.2, cex.lab=1.2))
with(pheno.egcut, points(Age, loyc, pch=20, col="blue"))
abline(lm(loyc ~ Age, data=pheno.egcut), lwd=2, col="red")
```




```
with(pheno.egcut, boxplot(Age ~ loy, ylab="Age"))
```



3 TCGA data

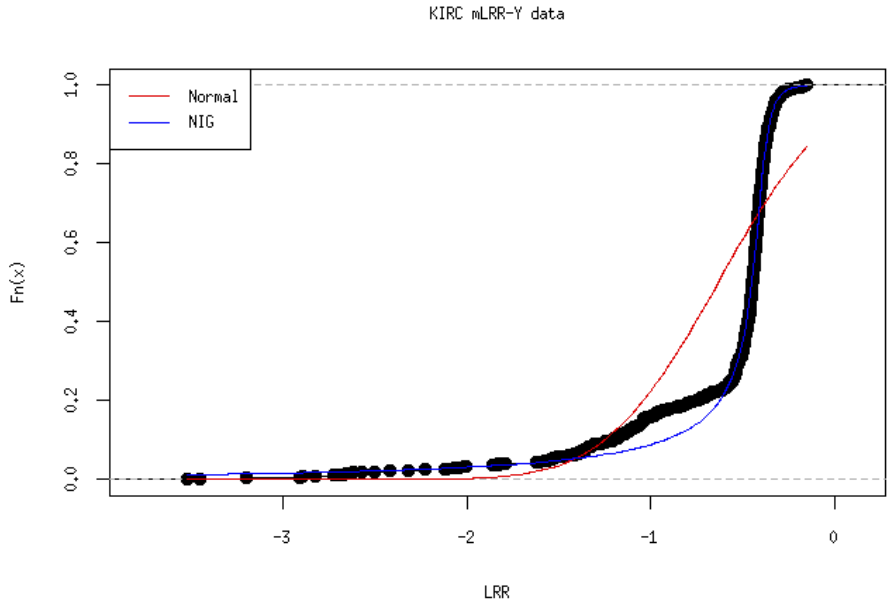
3.1 KIRC

Let us process the KIRC males raw data

```
path.kirc <- "/SYNCRW10125/PROJECTS/Aina_MSc/KIRC/rawData"
kirc <- madloy(path.kirc, trim=0.25, LRRCol=5, mc.cores=26)
```

Let us perform the calling. In that case, since data belongs to a more heterogeneous dataset, the NIG distribution shows even better fit than in the case of EGCUT data that corresponds to individuals from general population

```
plotNIG(kirc, "KIRC mLRR-Y data")
```




```
kirc.call <- getLOY(kirc)
kirc.call
```

```
##
## Object of class LOY
## ---------------------------
## Number of normal samples: 501
## Number of LOY: 162
## Number of XYY: 18
## Number of samples do not pass QC: 7
```

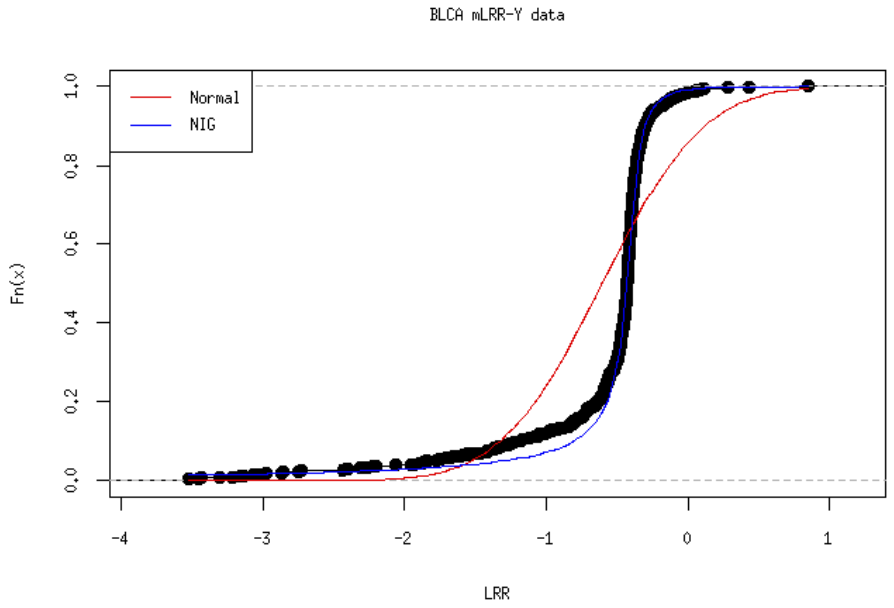
3.2 BLCA

Let us process the BLCA males raw data

```
path.blca <- "/SYNCRW10125/PROJECTS/Aina_MSc/BLCA/rawData"
blca <- madloy(path.blca, trim=0.25, LRRCol=5, mc.cores=26)
```

As in the case of KIRC data, the NIG distribution fits better mLRR-Y data than normal distribution

```
plotNIG(blca, "BLCA mLRR-Y data")
```



The calling procedure is then performed by executing

```
blca.call <- getLOY(blca)
blca.call
```

```
##
## Object of class LOY
## ---------------------------
## Number of normal samples: 488
## Number of LOY: 146
## Number of XYY: 36
## Number of samples do not pass QC: 5
```

3.3 Cancer association analysis

3.3.1 KIRC data

Let's us perform association analysis between tumoral stage (normal/tumor sample) and the LOY status using different strategies. This can be peformed for KIRC dataset by

```
tumor <- rep(0, length(kirc.call$class))
tumor[grep("01", names(kirc.call$class))] <- 1


pheno.kirc <- data.frame(id=substr(names(kirc.call$class), 6, 12),
                tumor=tumor,
                loy=kirc.call$class,
                loyc=kirc.call$data)
rownames(pheno.kirc) <- substr(names(kirc.call$data), 1, 16)

tt <- getThreshold(pheno.kirc$loyc)
pheno.kirc$loyt <- relevel(cut(pheno.kirc$loyc, c(-Inf, tt, Inf)),2)

mod.kirc <- glm(tumor ~ loy, data=pheno.kirc, family=binomial)
mod.kirc.c <- glm(tumor ~ loyc, data=pheno.kirc, family=binomial)
mod.kirc.t <- glm(tumor ~ loyt, data=pheno.kirc, family=binomial)
```

The results using the called LOY are:

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.5958688 0.09334856 -6.383267 1.733490e-10
## loyLOY       3.3171642 0.33954574  9.769418 1.523250e-22
## loyXYY       1.2890160 0.50863931  2.534244 1.126903e-02
```

The pvalue of LRT test using the threshold approach used in Fosbert et al. 2014 and Dumansky et al. 2015 is:

```
##                     Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)       -0.5148377 0.08959982 -5.745967 9.139696e-09
## loyt(-Inf,-0.217]  3.6859227 0.42614560  8.649445 5.175070e-18
```

The results using LOY as quantitative variable are:

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.3469535 0.08632255 -4.019268 5.837911e-05
## loyc        -2.8975750 0.37595388 -7.707262 1.285459e-14
```

3.3.2 BLCA data

```
tumor <- rep(0, length(blca.call$class))
tumor[grep("01", names(blca.call$class))] <- 1

pheno.blca<- data.frame(id=substr(names(blca.call$class), 6, 12),
                tumor=tumor,
                loy=blca.call$class,
                loyc=blca.call$data)
rownames(pheno.blca) <- substr(names(blca.call$data), 1, 16)

tt <- getThreshold(pheno.kirc$loyc)
pheno.blca$loyt <- relevel(cut(pheno.blca$loyc, c(-Inf, tt, Inf)),2)

mod.blca <- glm(tumor ~ loy, data=pheno.blca, family=binomial)
mod.blca.c <- glm(tumor ~ loyc, data=pheno.blca, family=binomial)
mod.blca.t <- glm(tumor ~ loyt, data=pheno.blca, family=binomial)
```

The results using the called LOY are:

```
##               Estimate  Std. Error     z value     Pr(>|z|)
## (Intercept) -0.5212969   0.1023973 -5.09092685 3.563175e-07
## loyLOY       1.7521772   0.2227954  7.86451278 3.705368e-15
## loyXYY      18.0873654 659.3633960  0.02743156 9.781155e-01
```

The results using threshold method are:

```
##                     Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)       -0.2593725  0.0927345 -2.796936 5.158970e-03
## loyt(-Inf,-0.217]  1.7222069  0.2541270  6.776953 1.227364e-11
```

The results using LOY as quantitative variable are:

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.1171188 0.08752611 -1.338102 1.808633e-01
## loyc        -1.0903672 0.22273966 -4.895254 9.817879e-07
```

3.4 Age association analysis

```
library(RTCGA.clinical)
clin <- KIRC.clinical

age <- clin$"patient.age_at_initial_pathologic_diagnosis"

b10 = as.character(lapply(KIRC.clinical[,621], function(v) {
  if (is.character(v)) return(toupper(v))
  else return(v)
}))

b01 = as.character(lapply(KIRC.clinical[,1760], function(v) {
  if (is.character(v)) return(toupper(v))
  else return(v)
}))

a = data.frame(barcode = b01, age = age)
b = data.frame(barcode = b10, age = age)
ab = rbind(a,b)

cancer = data.frame(row.names=ab$barcode, barcode=ab$barcode, age=ab$a

ids = intersect(rownames(cancer),rownames(pheno.kirc))
cancer.ok = cancer[ids,]
pheno.ok = pheno.kirc[ids,]

pheno.ok$y <- as.numeric(as.character(cancer.ok$age))
summary(lm(y ~  loyc + tumor, data=pheno.ok))
```

```
##
## Call:
## lm(formula = y ~ loyc + tumor, data = pheno.ok)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.060  -9.212  -0.116   8.856  28.647
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.1397     0.6659  88.809  < 2e-16 ***
## loyc         -3.7191     1.0041  -3.704  0.00023 ***
## tumor        -1.3526     0.9891  -1.368  0.17193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.92 on 659 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.0204, Adjusted R-squared:  0.01743
## F-statistic: 6.863 on 2 and 659 DF,  p-value: 0.001122
```

```
summary(lm(y ~  loy + tumor, data=pheno.ok))
```

```
##
## Call:
## lm(formula = y ~ loy + tumor, data = pheno.ok)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.083  -9.083  -0.162   8.917  28.917
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.0833     0.6701  88.168  < 2e-16 ***
## loyLOY        4.4420     1.2436   3.572  0.00038 ***
## loyXYY        0.1423     2.8840   0.049  0.96067
## tumor        -1.9216     1.0663  -1.802  0.07197 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.94 on 658 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.01932,    Adjusted R-squared:  0.01485
## F-statistic: 4.321 on 3 and 658 DF,  p-value: 0.004982
```

```
clin <- BLCA.clinical

age <- clin$"patient.age_at_initial_pathologic_diagnosis"

b10 = as.character(lapply(BLCA.clinical[,897], function(v) {
  if (is.character(v)) return(toupper(v))
  else return(v)
}))

b01 = as.character(lapply(BLCA.clinical[,1709], function(v) {
  if (is.character(v)) return(toupper(v))
  else return(v)
}))

a = data.frame(barcode = b01, age = age)
b = data.frame(barcode = b10, age = age)
ab = rbind(a,b)

cancer = data.frame(row.names=ab$barcode, barcode=ab$barcode, age=ab$a

ids = intersect(rownames(cancer),rownames(pheno.blca))
cancer.ok = cancer[ids,]
pheno.ok = pheno.blca[ids,]

pheno.ok$y <- as.numeric(as.character(cancer.ok$age))
summary(lm(y ~  loyc + tumor, data=pheno.ok))
```

```
##
## Call:
## lm(formula = y ~ loyc + tumor, data = pheno.ok)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.235  -7.277  -0.165   7.771  22.208
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  67.2827     0.6327 106.335   <2e-16 ***
## loyc         -1.3196     0.7928  -1.665   0.0966 .
## tumor        -0.3347     0.9003  -0.372   0.7102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.57 on 578 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.004771,   Adjusted R-squared:  0.001327
## F-statistic: 1.385 on 2 and 578 DF,  p-value: 0.2511
```

```
summary(lm(y ~  loy + tumor, data=pheno.ok))
```

```
##
## Call:
## lm(formula = y ~ loy + tumor, data = pheno.ok)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.893  -6.893   0.195   8.107  22.107
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.8934     0.6393 104.634  < 2e-16 ***
## loyLOY        3.8738     1.0817   3.581 0.000371 ***
## loyXYY        0.7786     1.9183   0.406 0.684985
## tumor        -1.0887     0.9584  -1.136 0.256448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.48 on 577 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.02211,    Adjusted R-squared:  0.01703
## F-statistic: 4.349 on 3 and 577 DF,  p-value: 0.00483
```

4 Alzheimer's disease data

Dumansky et al (2016) found significant association between Alzheimer's disease and LOY in three independent cohorts (ULSAM, PIVUS30 and EADI1). They scored participants as affected with LOY or not by using the lower limit of the 99% confidence interval of the experimentally induced variation of the mLRRY distribution for each cohort separately by using the threshold method described in the Methods section. Here we show that our proposed methodology also validate their findings by analayzing the NIA cohort that is available at dbGAP under the accession number phs000168v2.

Let us load the pre-process and phenotypic data. Data has been pre-process using another script that is not illustrated here just to guarantee anonymity of subjects. We first select males samples and then apply the two main steps `madloy` and `getLOY`. R code can be obtained upon request.

```
load("/SYNCRW10125/PROJECTS/LOY/Alzheimer/phs000168v2.NIA/Alzheimer_LO
dbgap.id <-unlist(lapply(strsplit(names(nia.call$class), "\\."), funct
identical(dbgap.id, pheno.nia$sample)
```

```
## [1] TRUE
```

```
pheno.nia$loy <- relevel(as.factor(nia.call$class), 2)
pheno.nia$loyc <- nia.call$data
tt <- getThreshold(pheno.nia$loyc)
pheno.nia$loyt <- relevel(cut(pheno.nia$loyc, c(-Inf, tt, Inf)),2)

pheno.nia$AgeDxDem[pheno.nia$AgeDxDem==999] <- NA
pheno.nia$event <- NA
pheno.nia$event[!is.na(pheno.nia$AgeDxDem)] <- 1

table(pheno.nia$loy)
```

```
##
## normal    LOY    XYY
##    608     36      1
```

```
mod.nia <- coxph(Surv(AgeDxDem, event) ~ loy, data=pheno.nia)
```

```
## Warning in coxph(Surv(AgeDxDem, event) ~ loy, data = pheno.nia): X
## deemed to be singular; variable 2
```

```
mod.nia.t <- coxph(Surv(AgeDxDem, event) ~ loyt, data=pheno.nia)
mod.nia.c <- coxph(Surv(AgeDxDem, event) ~ loyc, data=pheno.nia)
```

The results using calling LOY are:

```
## Call:
## coxph(formula = Surv(AgeDxDem, event) ~ loy, data = pheno.nia)
##
##   n= 278, number of events= 278
##    (367 observations deleted due to missingness)
##
##           coef exp(coef) se(coef)      z Pr(>|z|)
## loyLOY -0.6657    0.5139   0.2339 -2.846  0.00442 **
## loyXYY      NA        NA   0.0000     NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## loyLOY    0.5139      1.946    0.3249    0.8128
## loyXYY        NA         NA        NA        NA
##
## Concordance= 0.533  (se = 0.011 )
## Rsquare= 0.034   (max possible= 1 )
## Likelihood ratio test= 9.76  on 1 df,   p=0.001786
## Wald test            = 8.1  on 1 df,   p=0.004424
## Score (logrank) test = 8.4  on 1 df,   p=0.003758
```

```
##              coef exp(coef)  se(coef)         z   Pr(>|z|)
## loyLOY -0.6657498 0.5138881 0.2339061 -2.846227 0.00442407
## loyXYY         NA        NA 0.0000000        NA         NA
```

The results using mLRR as a quantitative variable are:

```
##           coef exp(coef)  se(coef)        z   Pr(>|z|)
## loyc 0.8771898  2.404134 0.3874427 2.264051 0.02357101
```

The results using threshold method are:

```
##                          coef exp(coef)  se(coef)         z    Pr(>
## loyt(-Inf,-0.0883] -0.4529163 0.6357714 0.1526103 -2.967796 0.00299
```

5 Transcriptomic analysis

Now, let us illustrate that samples having LOY are having transcriptomic effects in cancer data

5.1 KIRKC data

Let's load transcriptomic data from RTCGA package and get common samples and genes having more than 5 counts in more than 2 samples (recommended in DESeq2).

```
library(RTCGA.rnaseq)
counts <- KIRC.rnaseq
rownames(counts) <- substr(counts[,1], 1, 16)
counts <- t(ceiling(counts[,-1]))
filter <- apply(counts, 1, function(x) length(x[x>5])>=2)
counts.f <- counts[filter,]
pheno.kirc.c <- pheno.kirc[complete.cases(pheno.kirc),]
ii <- intersect(colnames(counts.f), rownames(pheno.kirc.c))
pheno.kirc <- pheno.kirc[ii,]
counts.kirc <- counts.f[,ii]
```

Now, let us peformed differential expression analysis by using voom method implemented in limma package. Surrogate variable analysis to remove unwanted variability is also performed with sva package:

```
fit.kirc <- de_voom(~ loy + tumor, ~ tumor, counts.kirc, pheno.kirc)
```

```
## Number of significant surrogate variables is:  13
## Iteration (out of 5 ):1  2  3  4  5
```

The top genes associated with LOY are:

```
topTable(fit.kirc, coef="loyLOY", adjust.method="fdr", sort.by="B", re
```

```
##                        logFC    AveExpr         t      P.Value    adj
## NCRNA00185|55410 -2.659642 -0.8218460 -25.44562 1.320251e-83 2.485
## TTTY4C|474150    -3.689755 -2.5403027 -24.70418 1.369057e-80 1.288
## UTY|7404         -2.046658  3.7192919 -24.26810 8.328791e-79 5.226
## TMSB4Y|9087      -2.109866  0.3672544 -24.08749 4.588824e-78 2.159
## USP9Y|8287       -2.284131  4.3752314 -24.05404 6.296308e-78 2.370
## CYorf15A|246126  -1.775502  2.6324642 -21.48585 2.860750e-67 8.975
## ZFY|7544         -1.591041  3.5773136 -21.32071 1.405479e-66 3.779
## EIF1AY|9086      -2.039787  3.7831495 -21.21412 3.929697e-66 9.247
## RPS4Y1|6192      -2.370463  7.0753540 -20.99865 3.145318e-65 6.578
## TTTY15|64595     -1.695636  2.2505865 -20.24695 4.515585e-62 8.500
##                        B
## NCRNA00185|55410 169.4591
## TTTY4C|474150    163.6545
## UTY|7404         165.4609
## TMSB4Y|9087      158.2131
## USP9Y|8287       164.2477
## CYorf15A|246126  138.6134
## ZFY|7544         138.5047
## EIF1AY|9086      137.5409
## RPS4Y1|6192      137.0570
## TTTY15|64595     126.7374
```

5.2 BLCA data

Let us load transcriptomic data from `RTCGA` package and get common samples and genes having more than 5 counts in more than 2 samples (recommended in `DESeq2`).

```
counts <- BLCA.rnaseq
rownames(counts) <- substr(counts[,1], 1, 16)
counts <- t(ceiling(counts[,-1]))
filter <- apply(counts, 1, function(x) length(x[x>5])>=2)
counts.f <- counts[filter,]
pheno.blca.c <- pheno.blca[complete.cases(pheno.blca),]
ii <- intersect(colnames(counts.f), rownames(pheno.blca.c))
pheno.blca <- pheno.blca[ii,]
counts.blca <- counts.f[,ii]
```

Now, let us peformed differential expression analysis by using `voom` method implemented in `limma` package. Surrogate variable analysis to remove unwanted variability is also performed with `sva` package:

```
fit.blca <- de_voom(~ loy + tumor, ~ tumor, counts.blca, pheno.blca)
```

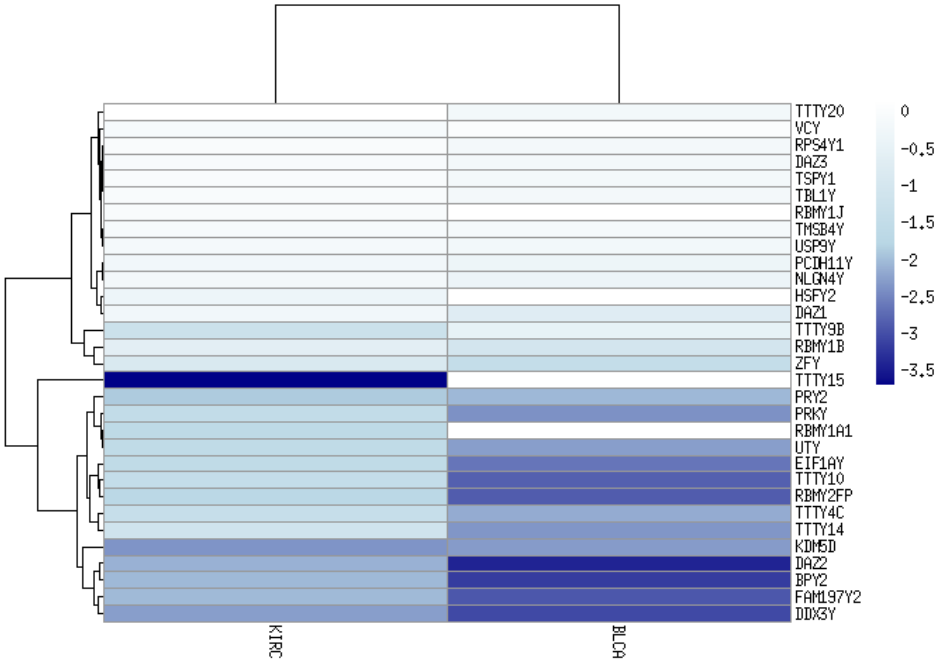
```
## Number of significant surrogate variables is:  9
## Iteration (out of 5 ):1  2  3  4  5
```

```
topTable(fit.blca, coef="loyLOY", adjust.method="fdr", sort.by="B", re
```

```
##                     logFC  AveExpr         t      P.Value    adj.P.
## UTY|7404        -3.183002 2.974611 -18.20832 2.354908e-50 4.484452e
## DDX3Y|8653      -2.639195 5.014783 -17.07262 4.427116e-46 4.215279e
## CYorf15A|246126 -3.164217 2.453245 -16.40572 1.433263e-43 9.097878e
## USP9Y|8287      -3.059661 3.336676 -16.07188 2.578963e-42 1.227780e
## TMSB4Y|9087     -3.420297 0.522574 -15.72130 5.341149e-41 2.034230e
## CYorf15B|84663  -3.281847 2.400365 -15.48432 4.129989e-40 1.310790e
## TTTY15|64595    -2.874175 1.754359 -15.45068 5.520292e-40 1.501756e
## KDM5D|8284      -2.838696 4.562966 -15.19566 4.967347e-39 1.182415e
## EIF1AY|9086     -2.927468 3.548773 -14.37618 5.594651e-36 1.183766e
## ZFY|7544        -2.281488 2.858952 -13.28196 5.939860e-32 1.131128e
##                        B
## UTY|7404        94.19555
## DDX3Y|8653      90.15330
## CYorf15A|246126 79.92424
## USP9Y|8287      79.01352
## TMSB4Y|9087     72.66963
## CYorf15B|84663  73.03027
## TTTY15|64595    71.55675
## KDM5D|8284      74.40259
## EIF1AY|9086     66.51378
## ZFY|7544        57.41446
```

5.3 Heatmap on chromosome Y genes

```
ensembl <- useMart("ensembl", dataset="hsapiens_gene_ensembl")
y <- getBM(attributes= "hgnc_symbol",
        filters = "chromosome_name",
        values = "Y",
        mart = ensembl)
genesY <- y$hgnc_symbol

kirc.y <- GenesRegion(fit.kirc, genesY)
temp1 <- data.frame(genes=kirc.y$intersect, KIRC=kirc.y$topTableRegior
blca.y <- GenesRegion(fit.blca, genesY)
temp2 <- data.frame(genes=blca.y$intersect, BLCA=blca.y$topTableRegior
temp3 <- merge(temp1, temp2, all.x=TRUE)
genesY.tumor <- temp3[,-1]
rownames(genesY.tumor) <- temp1[,1]


my_palette <- colorRampPalette(c("dark blue", "lightblue","white"))(n
pheatmap(genesY.tumor, color=my_palette)
```



6 Session information

```
## R version 3.3.0 (2016-05-03)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: CentOS release 6.7 (Final)
##
## locale:
##  [1] LC_CTYPE=es_ES.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=es_ES.UTF-8        LC_COLLATE=es_ES.UTF-8
##  [5] LC_MONETARY=es_ES.UTF-8    LC_MESSAGES=es_ES.UTF-8
##  [7] LC_PAPER=es_ES.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=es_ES.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     dat
## [8] methods   base
##
## other attached packages:
##  [1] RTCGA.rnaseq_20151101.4.0   RTCGA.clinical_20151101.4.0
##  [3] RTCGA_1.5.1                 RColorBrewer_1.1-2
##  [5] pheatmap_1.0.8              biomaRt_2.30.0
##  [7] sva_3.22.0                  genefilter_1.56.0
##  [9] mgcv_1.8-12                 nlme_3.1-127
## [11] limma_3.30.13               xtable_1.8-2
## [13] CNVassoc_2.2-1              survival_2.39-5
## [15] mclust_5.3                  mixdist_0.5-4
## [17] CNVassocData_1.0            lme4_1.1-12
## [19] Matrix_1.2-6                MADloy_0.9.7
## [21] GenomicRanges_1.26.4        GenomeInfoDb_1.10.3
## [23] IRanges_2.8.2               S4Vectors_0.12.2
## [25] BiocGenerics_0.20.0         data.table_1.10.4
## [27] knitr_1.16                  BiocStyle_2.2.1
##
## loaded via a namespace (and not attached):
##  [1] viridis_0.3.4               Biobase_2.34.0
##  [3] httr_1.2.1                  splines_3.3.0
##  [5] assertthat_0.2.0            yaml_2.1.14
##  [7] slam_0.1-34                 GeneralizedHyperbolic_0.8-1
##  [9] RSQLite_1.1-2               backports_1.0.5
## [11] lattice_0.20-33             glue_1.1.1
## [13] digest_0.6.12               XVector_0.14.1
## [15] rvest_0.3.2                 minqa_1.2.4
## [17] colorspace_1.3-2            htmltools_0.3.6
## [19] plyr_1.8.4                  XML_3.98-1.5
## [21] pkgconfig_2.0.1             zlibbioc_1.20.0
## [23] purrr_0.2.2                 scales_0.4.1
## [25] tibble_1.3.3                annotate_1.52.0
## [27] ggplot2_2.2.1               lazyeval_0.2.0
## [29] magrittr_1.5                memoise_1.0.0
## [31] evaluate_0.10.1             MASS_7.3-45
```

```
## [33] xml2_1.1.1                  ggthemes_3.3.0
## [35] tools_3.3.0                 stringr_1.2.0
## [37] munsell_0.4.3               AnnotationDbi_1.36.2
## [39] bindrcpp_0.2                survminer_0.2.4
## [41] rlang_0.1.1                 grid_3.3.0
## [43] RCurl_1.95-4.8              nloptr_1.0.4
## [45] bitops_1.0-6                rmarkdown_1.3
## [47] gtable_0.2.0                codetools_0.2-14
## [49] DBI_0.5-1                   R6_2.2.2
## [51] gridExtra_2.2.1             dplyr_0.7.1
## [53] bindr_0.1                   rprojroot_1.2
## [55] stringi_1.1.5               Rcpp_0.12.11
## [57] wordcloud_2.5
```

2.3 Supplementary Material 2

Simulation Studies: Robust estimation of mosaic loss of chromosome Y with genotype-array-intensity data

Juan R Gonzalez, Marcos Lopez, Pere Puig, ..., Tonu Esko, Luis A Perez-Jurado

Contents

1 Getting started
2 Quantitative traits
  2.1 Simulation
  2.2 Simulation results
3 Qualitative traits
  3.1 Simulation
  3.2 Simulation results
4 Session information

1 Getting started

Let us load required functions and packages

```
library(CNVassoc)
library(parallel)
source("R/fun_simul.R")
```

2 Quantitative traits

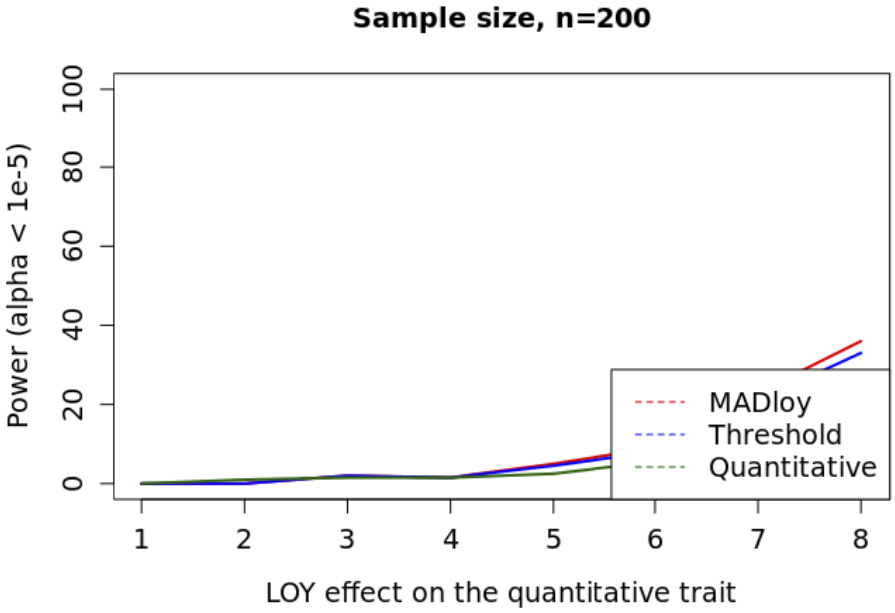
2.1 Simulation

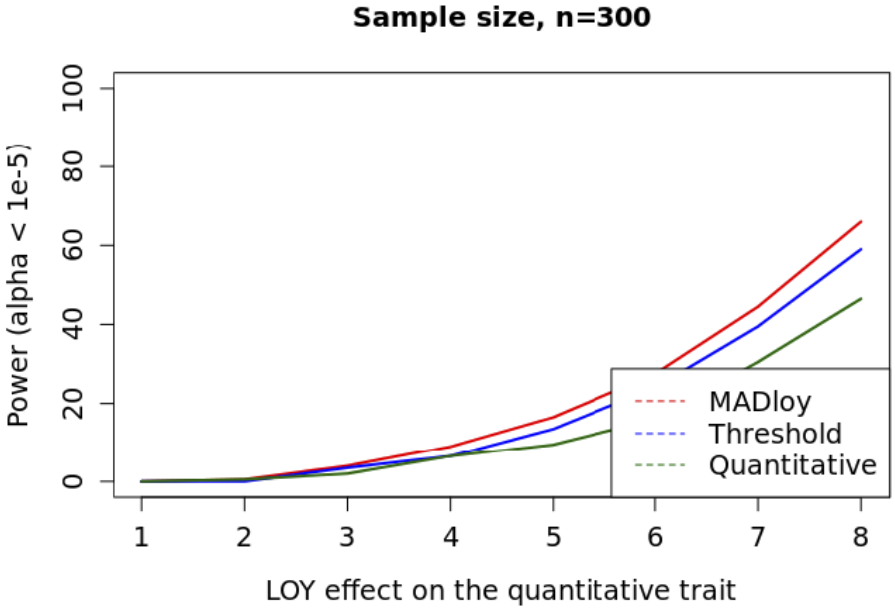
```
set.seed(123456)
ns <- c(200, 300, 500, 750, 1000, 1500)
effects <- c(1:8)
ans.eff.quant <- list()
for (i in 1:length(ns)) {
  ans.eff.quant[[i]] <- list()
  for(j in 1:length(effects)){
   ans.eff.quant[[i]][[j]] <- mclapply(1:200, simulation,
                                       type.trait="quant",
                                       n=ns[i], mu.surrog=c(0,-1.7),
                                       sd.surrog=c(0.04,0.9),
                                       w=c(0.9, 0.1),
                                       mu.y=50+c(0,effects[j]),
                                       sd.y=c(8,10), mc.cores=40)
  }
}
```

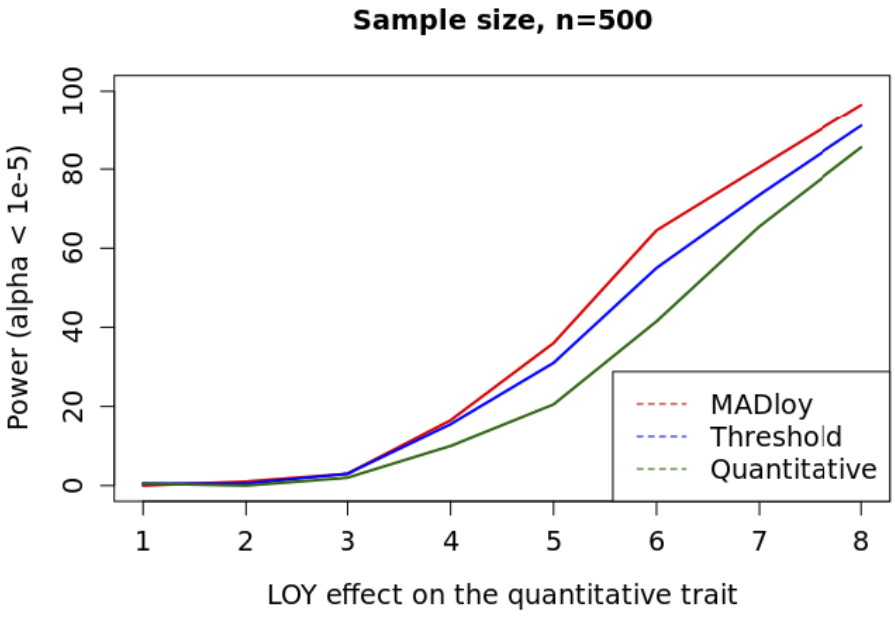
2.2 Simulation results

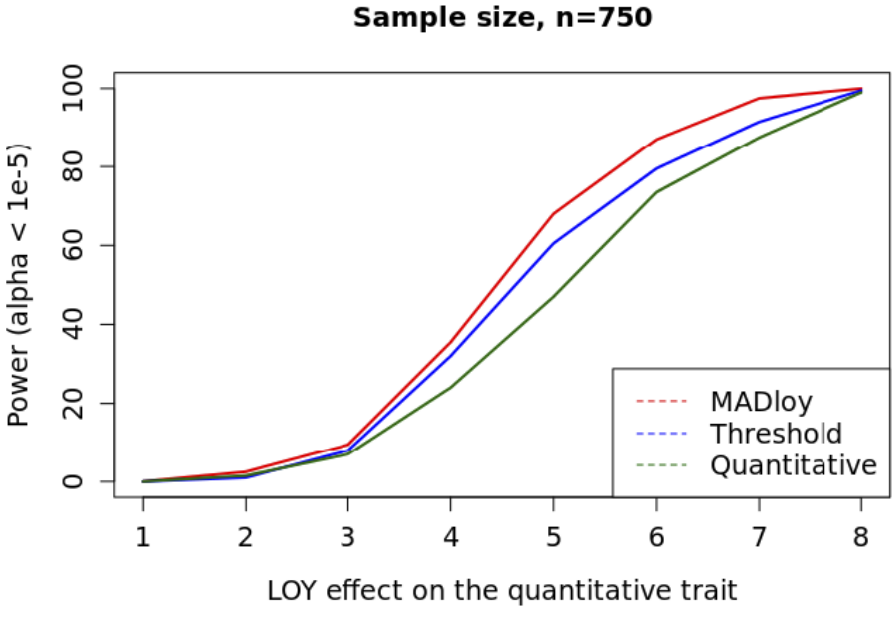
```
pval.eff.quant <- beta.eff.quant <- list()
for (i in 1:length(ns)){
  pval.eff.quant[[i]] <- beta.eff.quant[[i]] <- list()
  for (j in 1:length(effects)){
    temp <- do.call(rbind, lapply(ans.eff.quant[[i]][[j]], "[[", "pval
    pval.eff.quant[[i]][[j]] <- apply(temp, 2, function(x) mean(x<1e-5
    temp <- do.call(rbind, lapply(ans.eff.quant[[i]][[j]], "[[", "beta
    beta.eff.quant[[i]][[j]] <- apply(temp, 2, mean)
 }
}
```

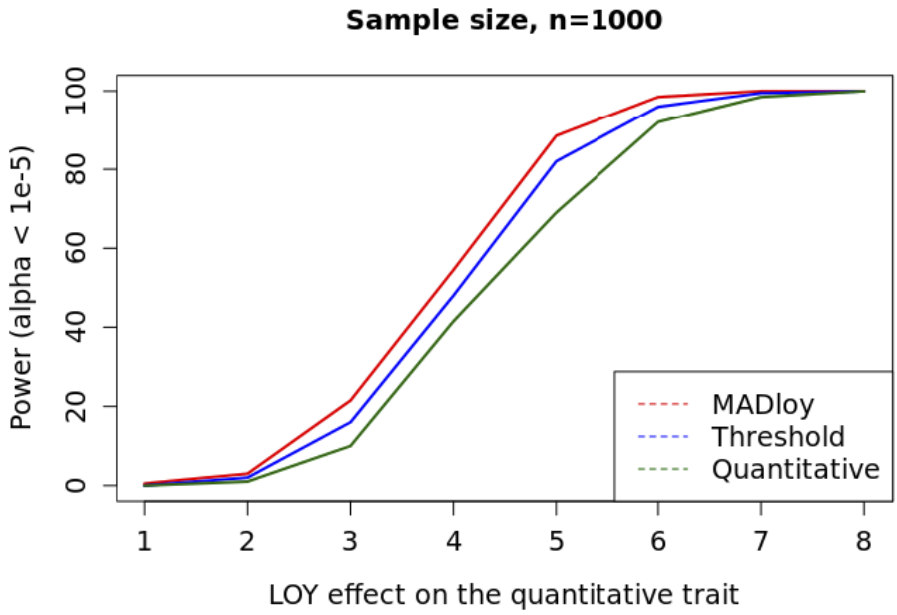
```
for (i in 1:length(ns)) {
dd <- do.call(rbind, pval.eff.quant[[i]])*100
plot(effects, dd[,1], type="n", xlab="LOY effect on the quantitative t
     ylab="Power (alpha < 1e-5)", ylim=c(0,100), cex.lab=1.2, cex.axis
title(paste0("Sample size, n=", ns[i]))
lines(effects, dd[,2], col="red", lwd=2)
lines(effects, dd[,3], col="blue", lwd=2)
lines(effects, dd[,4], col="darkgreen", lwd=2)
legend("bottomright", c("MADloy", "Threshold", "Quantitative"),
       col=c("red", "blue", "darkgreen"), lty=2, cex=1.2)
}
```

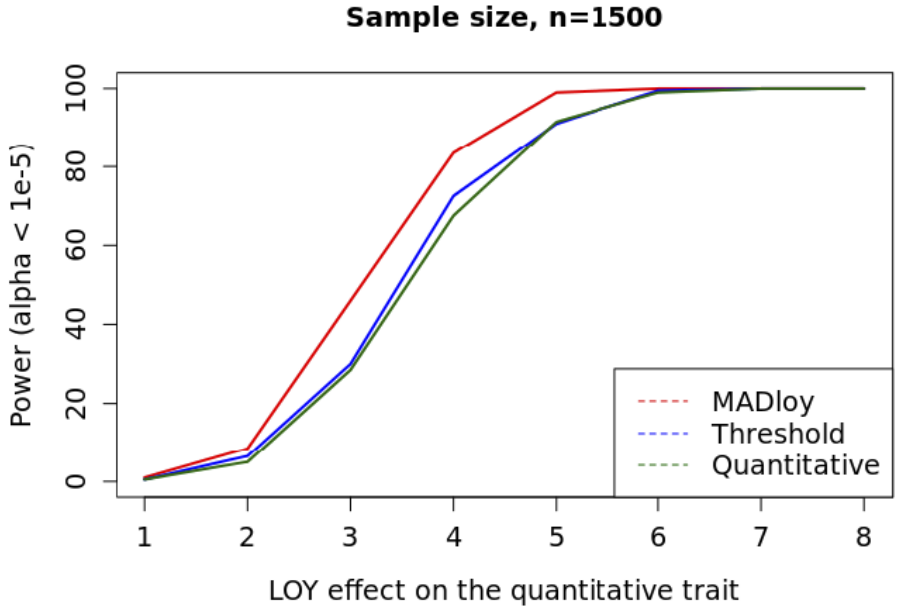

Sample size, n=200
Power (alpha < 1e-5)
0
20
40
60
80
100
1
2
3
4
5
6
7
8
LOY effect on the quantitative trait
MADloy
Threshold
Quantitative


Sample size, n=300
Power (alpha < 1e-5)
0
20
40
60
80
100
1
2
3
4
5
6
7
8
LOY effect on the quantitative trait
MADloy
Threshold
Quantitative


Sample size, n=500
Power (alpha < 1e-5)
0
20
40
60
80
100
1
2
3
4
5
6
7
8
LOY effect on the quantitative trait
MADloy
Threshold
Quantitative


Sample size, n=750
Power (alpha < 1e-5)
0
20
40
60
80
100
1
2
3
4
5
6
7
8
LOY effect on the quantitative trait
MADloy
Threshold
Quantitative





3 Qualitative traits

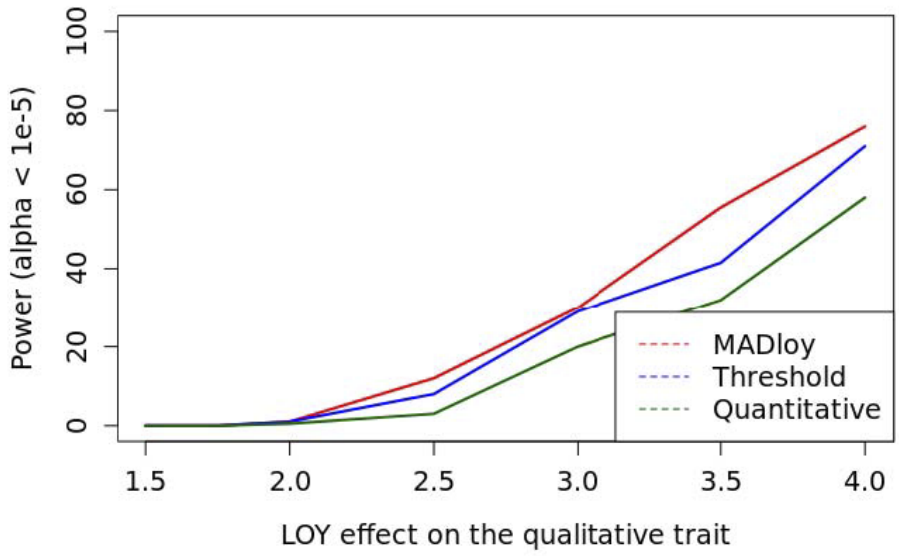
3.1 Simulation

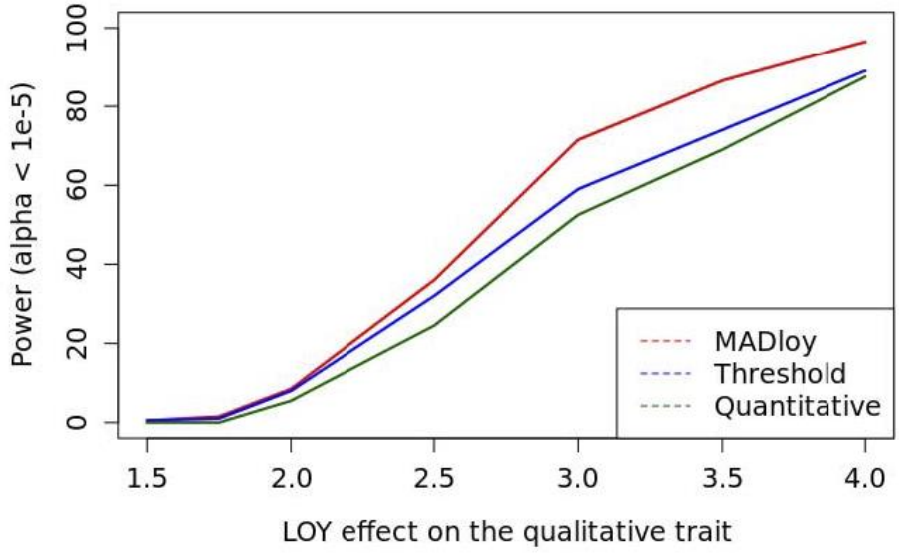
```
ns <- c(200, 300, 500, 750, 1000, 1500)
ors <- c(1.5, 1.75, 2, 2.5, 3, 3.5, 4)
ans.eff.quali <- list()
for (i in 1:length(ns)) {
  ans.eff.quali[[i]] <- list()
  for(j in 1:length(ors)){
   ans.eff.quali[[i]][[j]] <- mclapply(1:200, simulation,
                                       type.trait="quali",
                                       n0=ns[[i]], n1=ns[[i]],
                                       mu.surrog0=c(0,-0.40),
                                       sd.surrog0=c(0.04,0.24),
                                       mu.surrog1=c(0,-0.40),
                                       sd.surrog1=c(0.04,0.24),
                                       w0=c(0.9, 0.1),
                                       or=ors[j],
                                       mc.cores=40)
  }
}
```

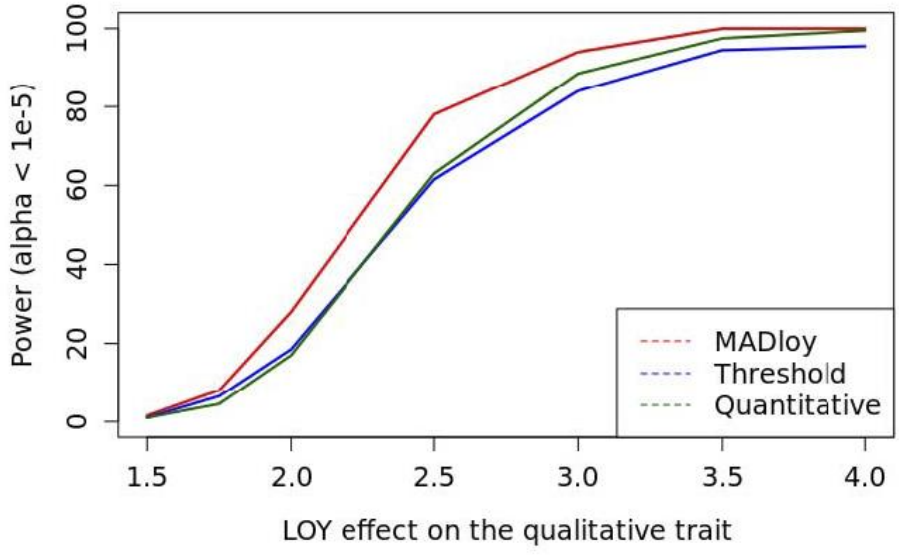
3.2 Simulation results

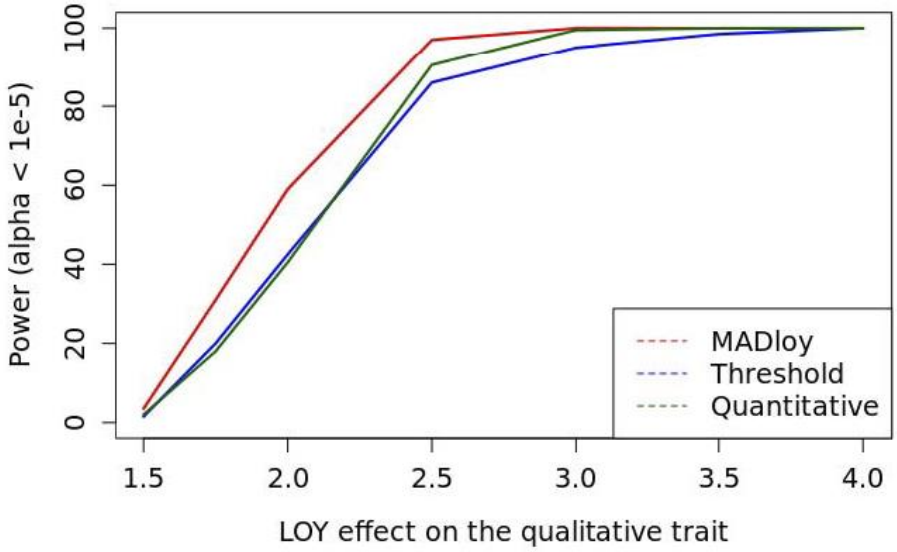
```
pval.eff.quali <- beta.eff.quali <- list()
for (i in 1:length(ns)){
  pval.eff.quali[[i]] <- beta.eff.quali[[i]] <- list()
  for (j in 1:length(ors)){
    o <- unlist(lapply(ans.eff.quali[[i]][[j]], function(x) inherits(x
    temp.i <- ans.eff.quali[[i]][[j]][!o]
    temp <- do.call(rbind, lapply(temp.i, "[[", "pval"))
    pval.eff.quali[[i]][[j]] <- apply(temp, 2, function(x) mean(x<1e-4
    temp <- do.call(rbind, lapply(temp.i, "[[", "beta"))
    beta.eff.quali[[i]][[j]] <- apply(temp, 2, mean)
 }
}
```

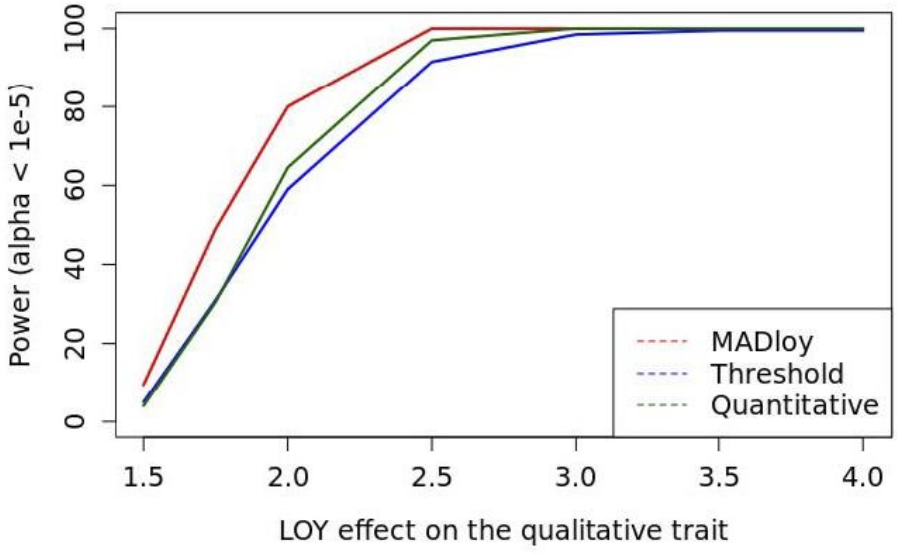
```
for (i in 1:length(ns)) {
dd <- do.call(rbind, pval.eff.quali[[i]])*100
plot(ors, dd[,1], type="n", xlab="LOY effect on the qualitative trait"
     ylab="Power (alpha < 1e-5)", ylim=c(0,100), cex.lab=1.2, cex.axis
lines(ors, dd[,2], col="red", lwd=2)
lines(ors, dd[,3], col="blue", lwd=2)
lines(ors, dd[,4], col="darkgreen", lwd=2)
legend("bottomright", c("MADloy", "Threshold", "Quantitative"),
       col=c("red", "blue", "darkgreen"), lty=2, cex=1.2)
}
```

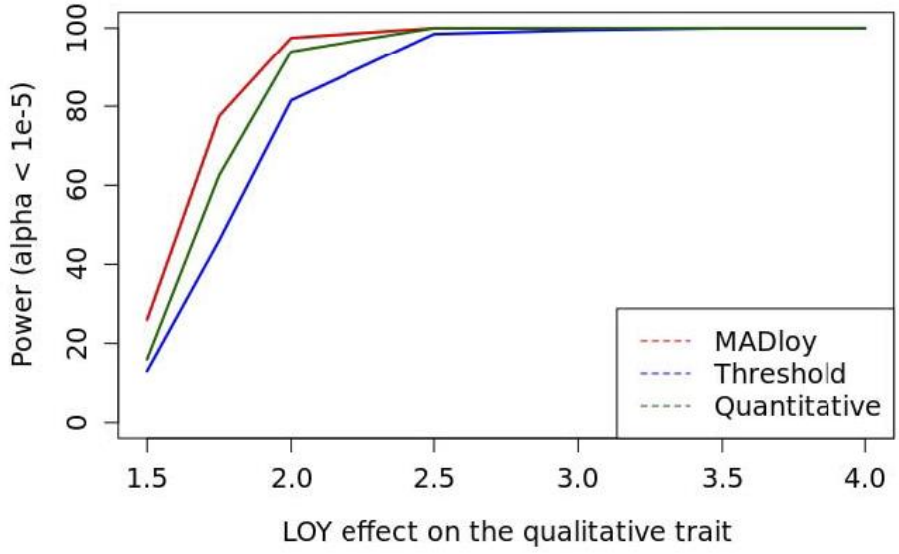



Power (alpha < 1e-5)
0
20
40
60
80
100
1.5
2.0
2.5
3.0
3.5
4.0
LOY effect on the qualitative trait
MADloy
Threshold
Quantitative


Power (alpha < 1e-5)
0
20
40
60
80
100
1.5
2.0
2.5
3.0
3.5
4.0
LOY effect on the qualitative trait
MADloy
Threshold
Quantitative


Power (alpha < 1e-5)
0
20
40
60
80
100
1.5
2.0
2.5
3.0
3.5
4.0
LOY effect on the qualitative trait
MADloy
Threshold
Quantitative


Power (alpha < 1e-5)
0
20
40
60
80
100
1.5
2.0
2.5
3.0
3.5
4.0
LOY effect on the qualitative trait
MADloy
Threshold
Quantitative



4 Session information

```
## R version 3.3.1 (2016-06-21)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.5 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C                LC_TIME=
##  [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=en_US.UTF-8     LC_MESSA
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                   LC_ADDRE
## [10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENT
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  met
##
## other attached packages:
## [1] CNVassoc_2.2     survival_2.41-3  mclust_5.3       mixdist_0.5-
## [6] knitr_1.16       BiocStyle_2.2.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.11     codetools_0.2-15 lattice_0.20-35  digest_0.6.
##  [6] grid_3.3.1       backports_1.1.0  magrittr_1.5     evaluate_0.
## [11] Matrix_1.2-10    rmarkdown_1.6    splines_3.3.1    tools_3.3.1
## [16] yaml_2.1.14      htmltools_0.3.6
```

2.4 Supplementary Material 3

Supplementary Material: assessing mLOY calling by Checking B deviation in PAR1 and PAR2 regioins

Juan R Gonzalez, Marcos Lopez, Pere Puig, Tonu Esko, Luis A Perez-Jurado

Contents

1 Getting started
2 EGCUT data
- 2.1 Check gender status
- 2.2 Preprocess data
- 2.3 LOY calling
- 2.4 Bdev assessing of the LOY classification
- 2.5 Bdev and ploidy

3 Session information

1 Getting started

Let us load required packages

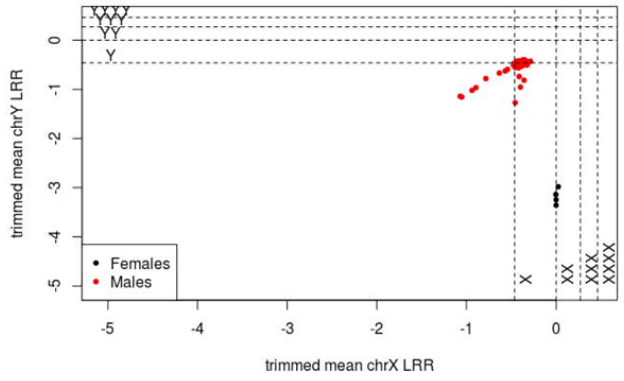
```
library(MADloy)
```

2 EGCUT data

2.1 Check gender status

```
path.egcut <- "/home/SHARED/DATA/EGCUT/Anon_Data/rawData_anon"
sex.egcut <- checkSex(path.egcut, mc.cores=15)
```

```
plot(sex.egcut)
```




```
egcut.males <- sex.egcut$par$files[sex.egcut$class=="MALE"]
```

2.2 Preprocess data

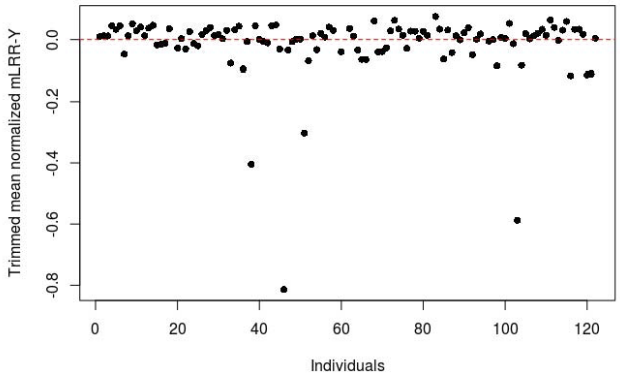
Let us preprocess raw data files to get LRR information in the mLRR-Y region and the autosomes

```
egcut <- madloy(egcut.males, trim=0.10, mc.cores=15)
egcut
```

```
## Object of class MADloy
## ----------------------------
## Number of processed samples: 122
## Target region: chrY:6671498-22919969
## Reference region(s): Autosomal chromosomes
## Offset (median LRR value in msY): -0.54
```

This figure illustrate the differences with regard to the mLRR-Y among individuals. Those having values around 0 are considered to be normal.

```
plot(egcut, ylim=c(-2, 1))
```

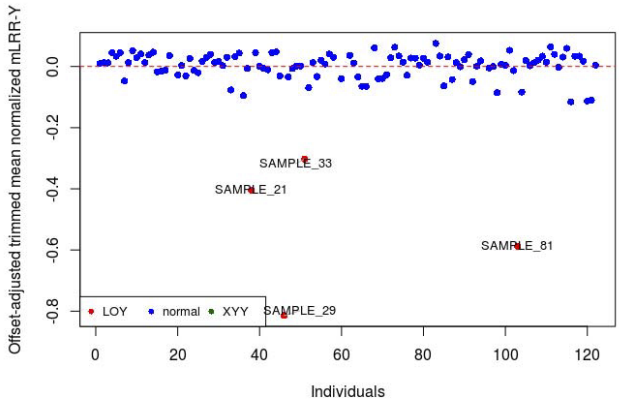


2.3 LOY calling

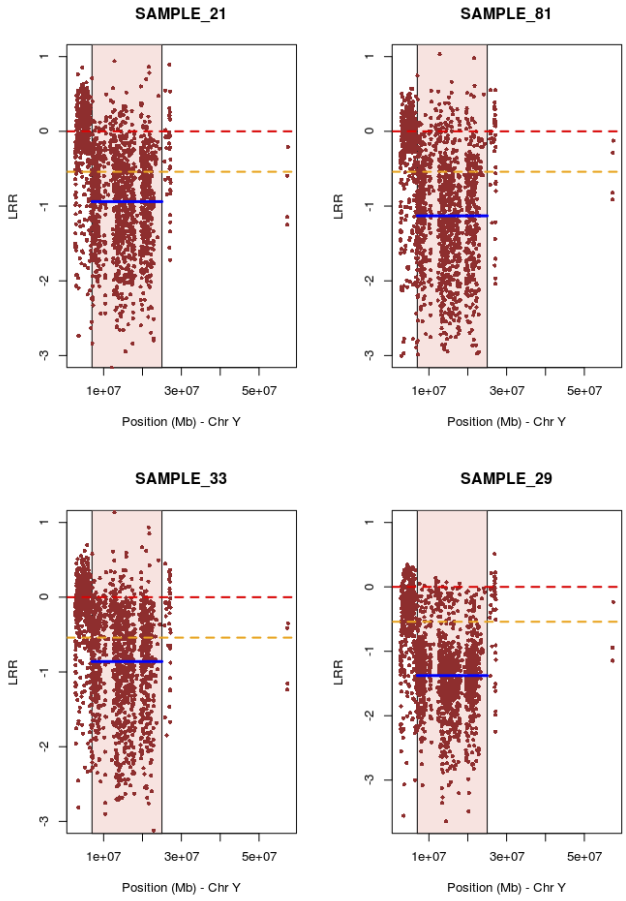
The calling is performed by using `getLOY` function. Notice that by default the `offset` argument of this function corresponds to the median value of mLRR-Y in all individuals. The `pval.sig` argument that is used to control the false discovery rate. By default it is based on Bonferroni correction that is set by default. In addition to the calling, those samples with LRR values with twice the mean standard deviation will be discarded due to their quality.

```
egcut.call <- getLOY(egcut)
```

```
plot(egcut.call, ylim=c(-2, 1), print.labels=TRUE)
```




```
par(mfrow=c(2,2))
plotIndLRR(egcut, sample="SAMPLE_21", ylim=c(-3,1))
plotIndLRR(egcut, sample="SAMPLE_81", ylim=c(-3,1))
plotIndLRR(egcut, sample="SAMPLE_33", ylim=c(-3,1))
plotIndLRR(egcut, sample="SAMPLE_29")
```



2.4 Bdev assessing of the LOY classification

In order to check the calling obtained by getLOY, we will check the B deviation in the PAR1 and PAR2 regions shared with chromosome X with the checkBdev function. The samples checked are those called by getLOY and the ones that does not pass the bonferroni correction but have a p-value lower than 0.05.

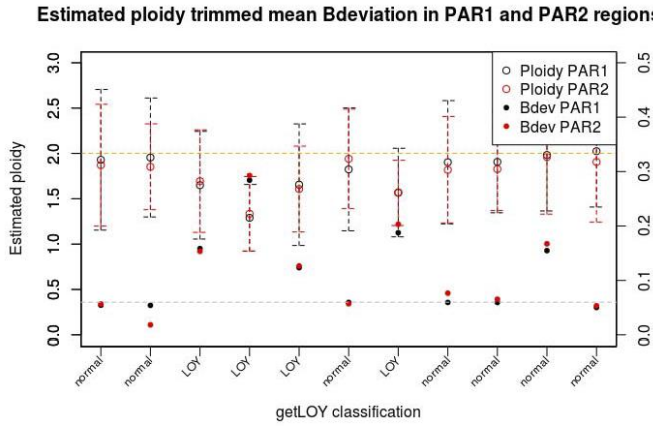
```
## Object of class MADloyBdev
## ---------------------------
##                   BdevClassification
## LRRClassification LOY normal
##            LOY      4      8
##            normal   1      6
```

```
knitr::kable(egcut.Bdev$class, caption="checkBdev results")
```

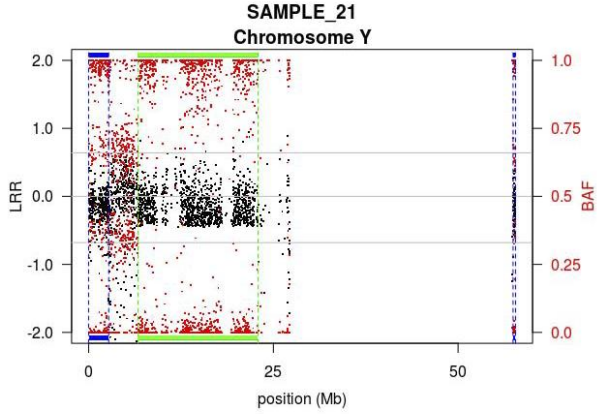
checkBdev results

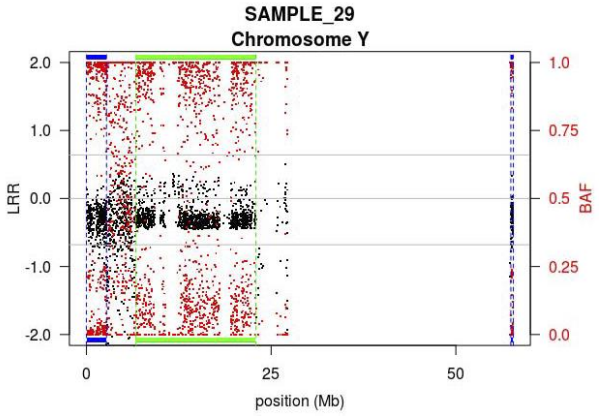
	orig	LRRPAR1	LRRPAR2	BdevPAR1	BdevPAR2	class	LRRCellPAR1	LRRCellPAR2	BdevCellPAR1	B
SAMPLE_17.txt	normal	-0.0238605	-0.0443736	0.054	0.057	normal	7.031564	12.878717	0.00	
SAMPLE_1.txt	normal	-0.0150924	-0.0508293	0.054	0.018	normal	4.476843	14.709767	0.00	
SAMPLE_21.txt	LOY	-0.1283232	-0.1105690	0.158	0.153	LOY	35.018640	30.565826	48.02	
SAMPLE_29.txt	LOY	-0.2823660	-0.2696150	0.284	0.283	LOY	71.005690	66.527577	72.45	
SAMPLE_33.txt	LOY	-0.1260569	-0.1454626	0.123	0.127	LOY	34.457570	39.206084	39.48	
SAMPLE_77.txt	normal	-0.0614249	-0.0197894	0.060	0.057	normal	17.604032	5.849585	0.00	

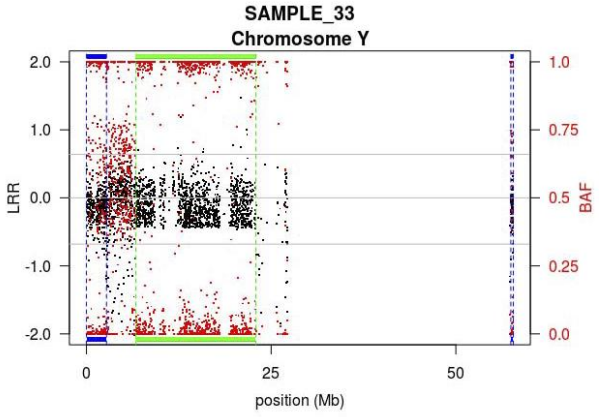
	orig	LRRPAR1	LRRPAR2	BdevPAR1	BdevPAR2	class	LRRCellPAR1	LRRCellPAR2	BdevCellPAR1	B
SAMPLE_81.txt	LOY	-0.1618274	-0.1634041	0.188	0.203	LOY	43.105088	43.475687	54.65	
SAMPLE_82.txt	normal	-0.0331216	-0.0633084	0.059	0.077	normal	9.683679	18.118889	0.00	
SAMPLE_84.txt	normal	-0.0318345	-0.0598541	0.059	0.065	normal	9.325910	17.173759	0.00	
SAMPLE_88.txt	normal	-0.0058108	-0.0129917	0.155	0.167	LOY	1.735673	3.859781	47.33	
SAMPLE_89.txt	normal	0.0081669	-0.0322002	0.050	0.053	normal	2.474251	9.430471	0.00	

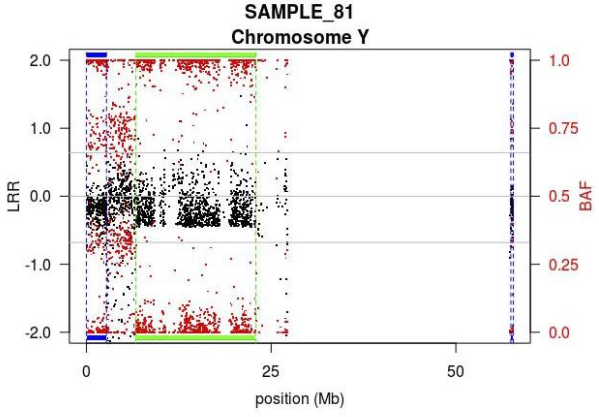


These can be visualized with the `plotIndSNP` function that draws the chromosome Y and highlights the mLRR-Y region (green) and PAR regions (blue)

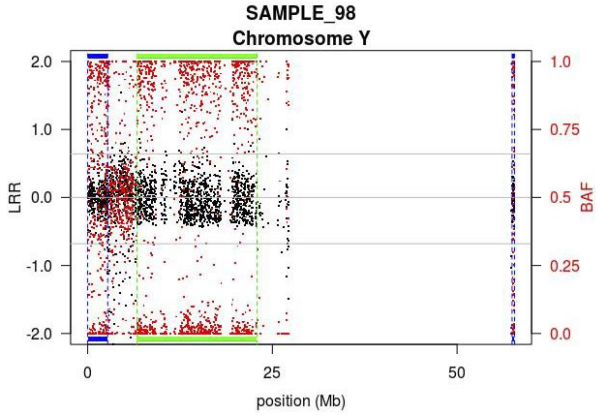


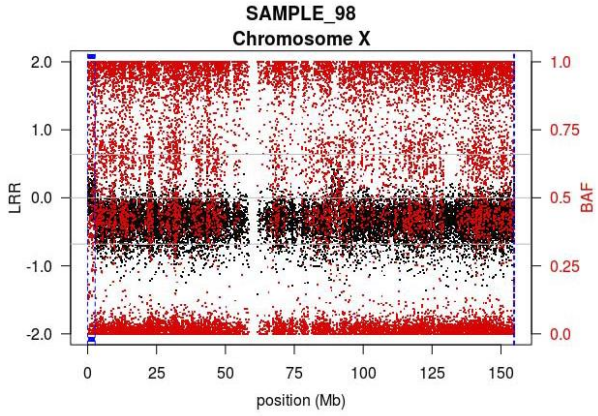

SAMPLE_29
Chromosome Y
LRR
BAF
2.0
1.0
0.0
-1.0
-2.0
1.0
0.75
0.5
0.25
0.0
0
25
50
position (Mb)


SAMPLE_33
Chromosome Y
LRR
BAF
2.0
1.0
0.0
-1.0
-2.0
1.0
0.75
0.5
0.25
0.0
0
25
50
position (Mb)


SAMPLE_81
Chromosome Y
LRR
BAF
2.0
1.0
0.0
-1.0
-2.0
1.0
0.75
0.5
0.25
0.0
0
25
50
position (Mb)

As can be seen, the SAMPLE_98 that has an altered B deviation value but has not called seems to be a mosaic Klinefelter. In order to check the chromosome X, there is an implemented function similar to the previous one that draws the chromosome called `plotIndSNPX`. The following generated plots show that part of the cells have a chromosome X gain, visible as an increased LRR values in PAR regions of chromosome Y and X and the B Allele Frequency split present in the whole chromosome X.



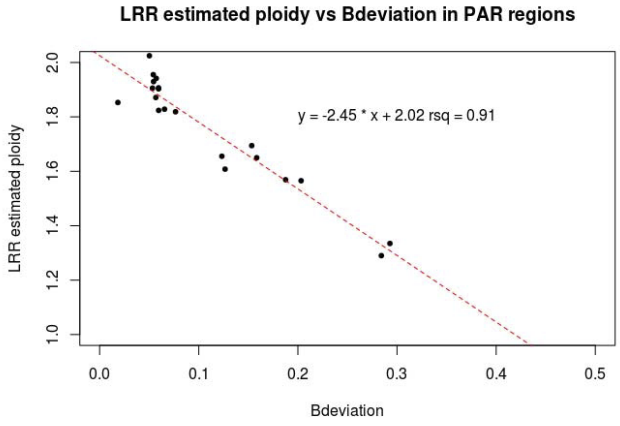


2.5 Bdev and ploidy

To estimate the correlation between B deviation and Log R Ratio measure, we removed SAMPLE_98 and estimated the ploidy of the trimmed mean LRR of the PAR regions in order to have normal-distributed values of ploidy. The formula used to estimate the ploidy is:

$$Ploidy = 2 * exp(1.5 * LRR)$$

There is a high correlation between the estimated ploidy of the PAR1 and PAR2 regions, median B deviation values.



3 Session information

```
## R version 3.4.1 (2017-06-30)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.3 LTS
##
## Matrix products: default
## BLAS: /usr/lib/libblas/libblas.so.3.6.0
## LAPACK: /usr/lib/lapack/liblapack.so.3.6.0
##
## locale:
##  [1] LC_CTYPE=es_ES.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=es_ES.UTF-8        LC_COLLATE=es_ES.UTF-8
##  [5] LC_MONETARY=es_ES.UTF-8    LC_MESSAGES=es_ES.UTF-8
##  [7] LC_PAPER=es_ES.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=es_ES.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats4    methods   stats     graphics  grDevices utils
## [8] datasets  base
##
## other attached packages:
## [1] MADloy_0.9.7         GenomicRanges_1.28.4 GenomeInfoDb_1.12.2
## [4] IRanges_2.10.2       S4Vectors_0.14.3     BiocGenerics_0.22.0
## [7] data.table_1.10.4    knitr_1.16           BiocStyle_2.4.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.12                XVector_0.16.0
##  [3] magrittr_1.5                zlibbioc_1.22.0
##  [5] wordcloud_2.5               stringr_1.2.0
##  [7] GeneralizedHyperbolic_0.8-1 tools_3.4.1
##  [9] htmltools_0.3.6             yaml_2.1.14
## [11] rprojroot_1.2               digest_0.6.12
## [13] GenomeInfoDbData_0.99.0     RColorBrewer_1.1-2
## [15] codetools_0.2-15            bitops_1.0-6
## [17] RCurl_1.95-4.8              slam_0.1-40
## [19] evaluate_0.10.1             rmarkdown_1.6
## [21] stringi_1.1.5               compiler_3.4.1
## [23] backports_1.1.0
```

3 Supplementary Material CHAPTER 3

Supplementary Table 1

Fiber-FISH and parent-of-origin data of 17 families. For individuals screened with the secondary probe set, detected inversion type is indicated.

Supplementary Figure 1

Circos v0.64 [46]; Paralogous relations among the four LCRs of chromosome 22q11.2. Numbers indicate hg38 chromosome 22 coordinates in Mb. Black bars on the outer circle delineate conventional coordinates of LCR22A, -B, -C, and –D. Blue rectangles in LCR22A depict gaps in the reference sequence. Segmental duplicated sequences are shown in orange [2,47] and the unique sequences in grey. Connecting lines, color-coded by percent identity, show Paralogous relations between subunits within and between LCR22s

Supplementary Figure 2

Detailed InveRsion plot of the chromosomal region 22q11 using single nucleotide variation data from the European samples of the 1000G project. Segment sizes ranged from 0.2 to 3 Mb. Orange boxes highlight the location of the LCR blocks. Several putative inversions were found in the region with the strongest signal located between LCR22C and D.

Supplementary Figure 3

Variable sizes of LCR22D, marked by the white arrow, between individuals for both reference and inversion patterns.

Demaerel W, Hestand MS, Vergaelen E, Swillen A, López-Sánchez M, Pérez-Jurado LA, et al. RETRACTED: Nested Inversion Polymorphisms Predispose Chromosome 22q11.2 to Meiotic Rearrangements. Am J Hum Genet. 2017 Oct 5;101(4):616–22. DOI: 10.1016/j.ajhg.2017.09.002

4 Article: Detectable clonal mosaicism in blood as a biomarker of cancer risk in Fanconi anemia

Reina-Castillón J, Pujol R, López-Sánchez M, Rodríguez-Santiago B, Aza-Carmona M, González JR, et al. Detectable clonal mosaicism in blood as a biomarker of cancer risk in Fanconi anemia. Blood Adv. 2017 Jan 24;1(5):319–29. DOI: 10.1182/bloodadvances.2016000943